

NICHOLAS JACKSON

A COURSE IN
ABSTRACT ALGEBRA

DRAFT: JANUARY 31, 2017

To Abigail and Emilie

Preface

We may always depend upon it that algebra which cannot be translated into good English and sound common sense is bad algebra.

— William Kingdon Clifford
(1845–1879),
The Common Sense of the Exact Sciences
(1886) 21

Mathematics is written for mathematicians

— Nicolaus Copernicus (1473–1543),
preface to *De Revolutionibus Orbium*
Cælestium (1543)

Contents

	<i>Preface</i>	v
1	<i>Groups</i>	1
1.1	<i>Numbers</i>	1
1.2	<i>Matrices</i>	17
1.3	<i>Symmetries</i>	20
1.4	<i>Permutations</i>	24
2	<i>Subgroups</i>	39
2.1	<i>Groups within groups</i>	39
2.2	<i>Cosets and Lagrange's Theorem</i>	49
2.3	<i>Euler's Theorem and Fermat's Little Theorem</i>	62
3	<i>Normal subgroups</i>	69
3.1	<i>Cosets and conjugacy classes</i>	69
3.2	<i>Quotient groups</i>	80
3.A	<i>Simple groups</i>	91
4	<i>Homomorphisms</i>	103
4.1	<i>Structure-preserving maps</i>	103
4.2	<i>Kernels and images</i>	110
4.3	<i>The Isomorphism Theorems</i>	116
5	<i>Presentations</i>	133
5.1	<i>Free groups</i>	134
5.2	<i>Generators and relations</i>	144
5.3	<i>Finitely generated abelian groups</i>	165
5.A	<i>Coset enumeration</i>	178
5.B	<i>Transversals</i>	186
5.C	<i>Triangles, braids and reflections</i>	197

6	<i>Actions</i>	205
6.1	<i>Symmetries and transformations</i>	205
6.2	<i>Orbits and stabilisers</i>	213
6.3	<i>Counting</i>	220
7	<i>Finite groups</i>	233
7.1	<i>Sylow's Theorems</i>	233
7.2	<i>Series of subgroups</i>	244
7.3	<i>Soluble and nilpotent groups</i>	255
7.4	<i>Semidirect products</i>	272
7.5	<i>Extensions</i>	277
7.A	<i>Classification of small finite groups</i>	291
8	<i>Rings</i>	317
8.1	<i>Numbers</i>	317
8.2	<i>Matrices</i>	326
8.3	<i>Polynomials</i>	328
8.4	<i>Fields</i>	332
8.A	<i>Modules and representations</i>	338
9	<i>Ideals</i>	343
9.1	<i>Subrings</i>	343
9.2	<i>Homomorphisms and ideals</i>	351
9.3	<i>Quotient rings</i>	359
9.4	<i>Prime ideals and maximal ideals</i>	368
10	<i>Domains</i>	377
10.1	<i>Euclidean domains</i>	377
10.2	<i>Divisors, primes and irreducible elements</i>	385
10.3	<i>Principal ideal domains</i>	391
10.4	<i>Unique factorisation domains</i>	395
10.A	<i>Quadratic integer rings</i>	408
11	<i>Polynomials</i>	421
11.1	<i>Irreducible polynomials</i>	421
11.2	<i>Field extensions</i>	425
11.3	<i>Finite fields</i>	439
11.4	<i>Field automorphisms and the Galois group</i>	442
11.5	<i>The Galois Correspondence</i>	447

11.6 Solving equations by radicals	464
11.A Geometric constructions	474
A Background	481
A.1 Sets	481
A.2 Functions	487
A.3 Counting and infinity	492
A.4 Relations	496
A.5 Number theory	499
A.6 Real analysis	502
A.7 Linear algebra	503
Index	505

However, there is a pleasure in recognizing old things from a new point of view. Also, there are problems for which the new point of view offers a distinct advantage.

— Richard Feynman (1918–1988),
Space-time approach to non-relativistic quantum mechanics, Reviews of Modern Physics 20 (1948) 367–387

1 Groups

OUR AIM IN THE STUDY OF ABSTRACT ALGEBRA is to consider familiar algebraic or numerical systems such as the integers or the real numbers, and distil from them certain sensible, universal properties. We then ask the question “what other things satisfy some or all of these properties?” and see if the answers give us any insight into either the universe around us, or mathematics itself.

In practice, this has been an astonishingly rich approach, yielding valuable insights not only into almost all branches of pure and applied mathematics, but large swathes of physics and chemistry as well.

In this chapter, we begin our study of **groups**: sets equipped with a single binary operation, satisfying certain basic criteria (associativity, existence of a distinguished identity element, existence of inverses). We will study a few different scenarios in which this structure naturally arises (number systems, matrices, symmetry operations in plane and solid geometry, and permutations of finite or infinite sets) and the links between them.

1.1 Numbers

THERE ARE MANY DIFFERENT WAYS we could begin our study of abstract algebra, but perhaps as sensible a place as any is with the set $\mathbb{N}$ of **natural numbers**. These are the counting numbers, with which we represent and enumerate collections of discrete physical objects. It is the first number system that we learn about in primary school; in fact, the first number system that developed historically. More precisely, $\mathbb{N}$ consists of the positive integers:

$$\mathbb{N} = \{1, 2, 3, \dots\}$$

(By convention, we do not consider zero to be a natural number.) So, we begin by studying the set $\mathbb{N}$ of natural numbers, and their properties under the operation of addition.

The Tao begets One,
One begets Two,
Two begets Three,
Three begets all things.

— Lao Tzu, *Tao Te Ching* 42:1–4

Perhaps the first thing we notice is that, given two numbers $a, b \in \mathbb{N}$, it doesn't matter what order we add them in, since we get the same answer either way round:

$$a + b = b + a \quad (1.1)$$

This property is called **commutativity**, and we say that addition of natural numbers is **commutative**.

Given three numbers $a, b, c \in \mathbb{N}$ to be added together, we have a choice of which pair to add first: do we calculate $a + b$ and then add c to the result, or work out $b + c$ and then add the result to a ? Of course, as we all learned at an early age, it doesn't matter. That is,

$$(a + b) + c = a + (b + c). \quad (1.2)$$

This property of the addition of natural numbers is also a pretty fundamental one, and has a special name: **associativity**. We say that addition of natural numbers is **associative**.

These two properties, commutativity and associativity, are particularly important ones in the study of abstract algebra and between them will form two cornerstones of our attempts to construct and study generalised versions of our familiar number systems.

But first of all, let's look carefully at what we have so far. We have a set, in this case $\mathbb{N}$, and an operation, in this case ordinary addition, defined on that set. Addition is one of the best-known examples of a **binary operation**, which we now define formally.

Definition 1.1 A **binary operation** defined on a set S is a function $f: S \times S \rightarrow S$.

¹ Definition A.7, page 486.

Here $S \times S$ is the Cartesian product¹ of S with itself: the set consisting of all ordered pairs (a, b) of elements of S . In other words, a binary operation is a function which takes as input an ordered pair of elements of the chosen set S , and gives us in return a single element of S .

Casting addition of natural numbers in this new terminology, we can define a function

$$f: \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}; \quad (a, b) \mapsto (a + b)$$

which maps two natural numbers $a, b \in \mathbb{N}$ to their sum. The commutativity condition can then be formulated as

$$f(a, b) = f(b, a) \quad \text{for all } a, b \in \mathbb{N}$$

which is perhaps slightly less intuitive than the original statement. But associativity fares somewhat worse:

$$f(f(a, b), c) = f(a, f(b, c)) \quad \text{for all } a, b, c \in \mathbb{N}$$

With the original statement of associativity of addition, it was fairly easy to see what was going on: we can ignore parentheses when

calculating sums of three or more natural numbers. Formulating addition as a function $f(a, b)$ in this way, we gain certain formal advantages, but we lose the valuable intuitive advantage that our original notation $a + b$ gives us.

So, to get the best of both worlds, we adopt the following notational convention: if our function f is a binary operation (rather than just some other function defined on $S \times S$) we will usually represent it by some symbol placed between the function's input values (which we will often refer to as **arguments** or **operands**). That is, instead of writing $f(a, b)$ we write $a * b$. In fact, given a binary operation $*$, we will usually adopt the same notation for the function: $*$: $S \times S \rightarrow S$.

Definition 1.2 A set S equipped with a binary operation $*$: $S \times S \rightarrow S$ is called a **magma**.

Formalising the above discussion, we give the following two definitions:

Definition 1.3 A binary operation $*$: $S \times S \rightarrow S$ defined on a set S is said to be **commutative** if

$$a * b = b * a$$

for all $a, b \in S$.

Definition 1.4 A binary operation $*$: $S \times S \rightarrow S$ defined on a set S is said to be **associative** if

$$(a * b) * c = a * (b * c)$$

for all $a, b, c \in S$.

At this point it's worth noting that we already have a sophisticated enough structure with which we can do some interesting mathematics.

Definition 1.5 A **semigroup** is a set S equipped with an associative binary operation $*$: $S \times S \rightarrow S$. If $*$ is also commutative, then S is a **commutative** or **abelian** semigroup.

Semigroup theory is a particularly rich field of study, although a full treatment is beyond the scope of this book.

There is another obvious binary operation on $\mathbb{N}$ which we typically learn about shortly after our primary school teachers introduce us to addition: multiplication. This operation $\cdot$: $\mathbb{N} \times \mathbb{N} \rightarrow \mathbb{N}$ is also both commutative and associative, but is clearly different in certain ways to our original addition operation. In particular, the number 1 has a special multiplicative property: for any number $a \in \mathbb{N}$, we find that

$$1 \cdot a = a \cdot 1 = a.$$

That is, multiplication by 1 doesn't have any effect on the other number



Wikimedia Commons / Commemorative stamp, USSR (1983)

Muḥammad ibn Mūsā al-Khwārizmī (c.780–c.850) was a Persian astronomer and mathematician whose work had a profound influence on the development of western mathematics and science during the later mediæval period. His best-known work, *al-Kitāb al-mukhtaṣar fī ḥisāb al-jabr wal-muqābala* (*The Compendious Book on Calculation by Completion and Balancing*) describes general techniques for solving linear and quadratic equations of various types. Here, *al-jabr* (“completion”) is an operation whereby negative terms are eliminated from an equation by adding an appropriate positive quantity to both sides, while *wal-muqābala* (“balancing”) is a method for simplifying equations by subtracting repeated terms from both sides.

Translated into Latin as *Liber algebrae et almucabala* in 1145 by the English writer Robert of Chester, the term *al-jabr* became our word “algebra”, while *Algorizmi*, the Latinised form of Al-Khwarizmi's name, is the origin of the word “algorithm”.

His other major works include *Zij al-Sindhind* (*Astronomical Tables of Sindh and Hind*), a collection of astronomical and trigonometrical tables calculated by methods developed in India, and the *Kitāb Ṣūrat al-Arḍ* (*Book of the Description of the Earth*), a reworking of the *Geographia*, an atlas written by the Alexandrian mathematician and astronomer Claudius Ptolemy (c.100–c.170) in the middle of the second century.

involved. No other natural number apart from 1 has this property with respect to multiplication. Also, there is no natural number which has this property with respect to addition: there exists no $z \in \mathbb{N}$ such that for any other $a \in \mathbb{N}$ we have $z + a = a + z = a$. Of course, from our knowledge of elementary arithmetic, we know of an obvious candidate for such an element, which alas doesn't happen to be an element of the number system $\mathbb{N}$ under investigation.

This leads us to two observations: firstly, that the multiplicative structure of $\mathbb{N}$ is fundamentally different in at least one way from the additive structure of $\mathbb{N}$. And secondly, it might be useful to widen our horizons slightly to consider number systems which have one (or possibly more) of these special neutral elements.

We will return to the first of these observations, and investigate the interplay between additive and multiplicative structures in more detail later, when we study the theory of rings and fields. But now we will investigate this concept of a neutral element, and in order to do so we state the following definition.

Definition 1.6 Let S be a set equipped with a binary operation $*$. Then an element $e \in S$ is said to be an **identity** or **neutral element** with respect to $*$ if

$$e * a = a * e = a$$

for all $a \in S$.

So, let's now extend our number system $\mathbb{N}$ to include the additive identity element 0. Denote this new set $\mathbb{N} \cup \{0\}$ by $\mathbb{N}_0$.

Definition 1.7 A **monoid** is a semigroup $(S, *)$ which has an identity element. If the binary operation $*$ is commutative then we say S is a **commutative** or **abelian** monoid.

Monoids also yield a rich field of mathematical study, and in particular are relevant in the study of automata and formal languages in theoretical computer science.

In this book, however, we are primarily interested in certain specialised forms of these objects, and so we return to our investigation of number systems. Historically, this system $\mathbb{N}_0$ represented an important conceptual leap forward, a paradigm shift from the simple enumeration of discrete, physical objects, allowing the explicit labelling of nothing, the case where there are no things to count.²

The mathematical concept of zero has an interesting history, a full discussion of which is beyond the scope of this book. However, a fascinating and readable account may be found in the book *The Nothing That Is* by Robert and Ellen Kaplan.³

² What's red and invisible?

No tomatoes.

— Anonymous

³ R Kaplan and E Kaplan, *The Nothing That Is: A Natural History of Zero*, Penguin (2000).

But having got as far as the invention of zero, it's not much further a step to invent negative numbers.⁴ With a bit of thought, we can formulate perfectly reasonable questions that can't be answered in either $\mathbb{N}$ or $\mathbb{N}_0$, such as "what number, when added to 3, gives the answer 2?"

Attempts to answer such questions, where the need for a consistent answer is pitted against the apparent lack of a physical interpretation for the concept, led in this case to the introduction of negative numbers. This sort of paradigm shift occurs many times throughout the history of mathematics: we run up against a question which is unanswerable within our existing context, and then ask "But what if this question had an answer after all?" This process is often a slow and painful one, but ultimately leads to an expanded understanding of the subject at hand. It took somewhere in excess of a thousand years for the concept of negative numbers to fully catch on. The Greek mathematician Diophantus, writing in the third century AD, rejected negative solutions to linear or quadratic equations as absurd. Even as late as the 16th century, the Italian mathematician Girolamo Cardano (1501–1576) referred to such numbers as *fictæ*, or fictitious, although his Italian predecessor Leonardo of Pisa, better known as Fibonacci, had interpreted them in a financial context as a loss or debit. Meanwhile, the Indian mathematicians Brahmagupta (598–668) and Mahāvira (9th century) had made the necessary intuitive leap and developed rules for multiplying negative numbers (although even Mahāvira balked at considering their square roots). Adjoining negative numbers to $\mathbb{N}_0$ yields the set of **integers**

$$\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}.$$

Later on, we will examine how $\mathbb{Z}$ was extended to construct more sophisticated number systems, in particular the **rational numbers** $\mathbb{Q}$, the **real numbers** $\mathbb{R}$ and the **complex numbers** $\mathbb{C}$, but for the moment this will suffice.

The operation of addition can be extended to the negative integers in an obvious and consistent way, and we find that all of those tricky questions involving subtraction can now be solved. More importantly, it turns out that for any integer $n \in \mathbb{N}_0$ there is a unique negative integer $-n \in \mathbb{Z}$ such that

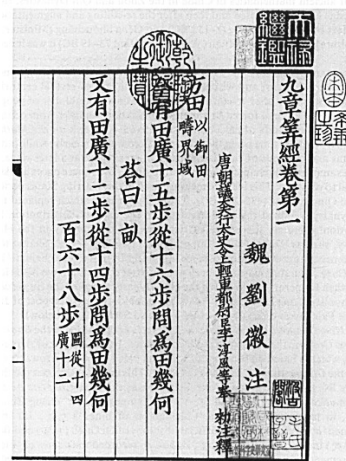
$$n + (-n) = 0 = (-n) + n. \quad (1.3)$$

But this also works for the negative integers themselves, as long as we define

$$-(-n) = n$$

for any integer $n \in \mathbb{Z}$. So, we now have a set $\mathbb{Z}$ equipped with

⁴ The earliest known treatment of negative numbers occurs in the ancient Chinese text *Jiuzhang suanshu* (*Nine Chapters on the Mathematical Art*), which dates from the Han dynasty (202BC–220AD) in which positive numbers are represented by black counting rods and negative numbers by red ones.



Wikimedia Commons

Figure 1.1: Page from *Nine Chapters on the Mathematical Art*



Wikimedia Commons / Unknown mediaeval artist

Leonardo of Pisa (c.1170–c.1250), commonly known as **Fibonacci**, is perhaps best known for the numerical sequence which bears his name, and which he discussed in his book *Liber Abaci* (1202) in the context of a simple model of population growth in rabbits. This sequence, which can be defined by the recurrence relation $F_{n+2} = F_{n+1} + F_n$ with $F_0 = 0$ and $F_1 = 1$, or by the formula

$$F_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n,$$

was known to Indian mathematicians as early as the 6th century, and manifests surprisingly often in the natural world: in the structure of artichokes and pinecones, and in the spiral arrangement of seeds in sunflowers.

Comparatively little is known of Leonardo himself, and the portrait above is believed to be a later invention not based on contemporary sources. The *Liber Abaci* notes that Leonardo was the son of a customs official named Guglielmo Bonaccio (the name Fibonacci is a contraction of *filii Bonacci*, or “son of Bonaccio”) and travelled with him to northern Africa. They spent some time in Bugia (now Bejaia in modern-day Algeria), where Leonardo recognised the usefulness of recent Arabic advances in mathematics.

After his return to Italy, Leonardo spent some time at the court of the Holy Roman Emperor Frederick II (1194–1250), who had a keen appreciation of mathematics and science. During this period he wrote other books, of which three survive: *Practica Geometriae* (1220), *Flos* (1225) and *Liber Quadratorum* (1225).

an associative and commutative binary operation $+: \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$, together with a designated identity element 0 and, for any number $n \in \mathbb{Z}$, an **inverse** element $-n$ satisfying (1.3). More generally, we have the following.

Definition 1.8 Let S be a set equipped with a binary operation $*$ and an identity element e . Then for any $a \in S$, an element a^{-1} is said to be an **inverse** of a if

$$a * a^{-1} = e = a^{-1} * a.$$

We are now able to define one of the most important structures in mathematics, whose properties we will study for much of this book.

Definition 1.9 A **group** $(G, *)$ consists of a set G together with a binary operation $*: G \times G \rightarrow G$ satisfying the following three criteria.

- G1** The binary operation $*$ is associative.
- G2** There exists an element $e \in G$ (the **identity** or **neutral element**) such that $e * g = g * e = g$ for all $g \in G$.
- G3** For each $g \in G$ there exists an element g^{-1} (the **inverse** of g) such that $g * g^{-1} = g^{-1} * g = e$.

Some texts include a fourth criterion:

- G0** The set G is **closed** under the action of $*$. That is, for any $g, h \in G$, it follows that $g * h \in G$ too.

However, in our case this is a direct consequence of the way we defined a binary operation: it is automatically the case that G is closed under the action of $*: G \times G \rightarrow G$.

When the group operation $*$ is obvious from context, we will often omit it, writing gh instead of $g * h$ for two elements $g, h \in G$. On other occasions, it may be notationally or conceptually more convenient to regard the group operation as a type of addition, rather than multiplication. In that case (especially if the group operation is commutative) we may choose to use additive notation, writing $g + h$ instead of $g * h$ or gh , and denoting the inverse of an element g by $-g$ rather than g^{-1} .

Although we’ve used e to denote the identity element of a group, this is by no means universal, and often we will use a different symbol, such as 0, 1, ι , or something else depending on the context.

Since a group is really a set with some extra structure, we can reuse most of the same concepts that we’re used to when dealing with sets, and in particular it’s often useful to consider a group’s cardinality:

Definition 1.10 The **order** of a group $G = (G, *)$ is the cardinality $|G|$ of its underlying set. A group is said to be **finite** (or **infinite**) if it has finite (or infinite) order.

Our motivating example $\mathbb{Z}$ is an infinite (more precisely, countably infinite) group, but shortly we will meet several finite examples.

As we saw earlier, many (but not all) well-known binary operations are commutative. This is certainly the case with the addition operation in $\mathbb{Z}$, which was the motivating example leading to our study of groups. So, on the premise that groups with commutative operations are important (which they are), we give them a special name:

Definition 1.11 An **abelian group** is a group $G = (G, *)$ whose operation $*$ is commutative. That is, $g * h = h * g$ for all $g, h \in G$.

The first few groups we will meet are all abelian, although in a short while we will study some examples of nonabelian groups as well.

Our first abelian example is a slight modification of $\mathbb{Z}$, but instead of taking the infinite set of integers, we take a finite subset, and instead of using the usual addition operation, we use modular arithmetic:

Example 1.12 (Cyclic groups) Let

$$\mathbb{Z}_n = \{0, \dots, n-1\}$$

be the set consisting of the first n non-negative integers, and let $+: \mathbb{Z}_n \times \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ be **addition modulo n** . That is, for any two $a, b \in \mathbb{Z}_n$ we define $a+b$ to be the remainder of the integer $a+b \in \mathbb{Z}$ after division by n .

This is the **cyclic group** of order n .

We can regard the elements of this group geometrically as n equally-spaced points around the circumference of a circle, and obtain $a+b$ by starting at point a and counting b positions clockwise round the circle to see which number we end up with. See Figure 1.2 for a geometric depiction of $5 + 9 = 2$ in $\mathbb{Z}_{12}$.

It's natural, when we meet a new mathematical construct, to ask what the simplest possible example of that construct is. In the case of groups, the following example answers this question.

Example 1.13 Let $G = \{0\}$ be the set consisting of a single element. There is only one possible binary operation that can be defined on this set, namely the one given by $0 * 0 = 0$. A routine verification shows that this operation satisfies all of the group axioms: 0 is the identity element, it's its own inverse, and the operation is trivially associative and commutative. This is the **trivial group**.

We could denote the trivial group as $\mathbb{Z}_1$, although nobody usually does: depending on the context we typically use 1 or 0 instead. Note that although we can define a binary operation of sorts (the empty operation) on the empty set $\emptyset$, we don't get a group structure because



Wikimedia Commons / Johan Gorbitz (1782–1853)

Abelian groups are named after the Norwegian mathematician **Niels Henrik Abel** (1802–1829), whose brilliant career (as well as his life) was cut tragically short by tuberculosis at the age of 26. At the age of 19 he proved, independently of his similarly tragic contemporary Évariste Galois (1811–1832), that the general quintic equation

$$ax^5 + bx^4 + cx^3 + dx^2 + ex + f = 0$$

cannot be solved by radicals. His monograph on elliptic functions was only discovered after his death, which occurred two days before the arrival of a letter appointing him to an academic post in Berlin.

In 2002, to commemorate his bicentenary (and approximately a century after the idea had originally been proposed) the Norwegian Academy of Sciences and Letters founded an annual prize in his honour, to recognise stellar achievement in mathematical research.

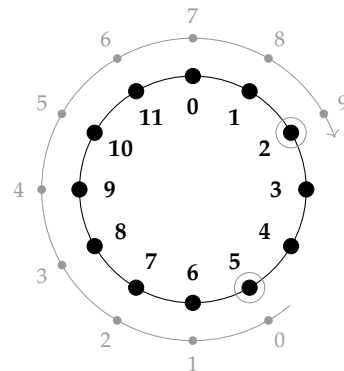


Figure 1.2: Addition in $\mathbb{Z}_{12}$. Here $5 + 9 = 14 \equiv 2 \pmod{12}$

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

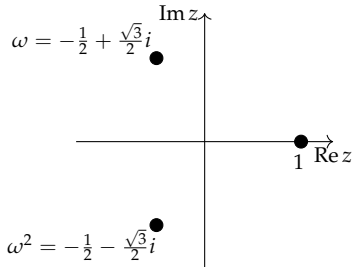
Table 1.1: Multiplication table for $\mathbb{Z}_4$ 

Figure 1.3: Cube roots of unity

axiom G2 requires the existence of at least one element: the identity. For a finite group, especially one of relatively small order, writing down the **multiplication table** is often the most effective way of determining the group structure. This is, as its name suggests, a table of all possible products of two elements of the group. Table 1.1 depicts the multiplication table (or, in this case, the **addition table**) for $\mathbb{Z}_4$.

In this book, we will adopt the convention that the product $a * b$ will be written in the cell where the a th row and b th column intersect. In the case where the group under investigation is abelian, its multiplication table will be symmetric about its leading diagonal, so this convention is only necessary when we study nonabelian groups.

Example 1.14 Let $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$ be a complex cube-root of unity, that is, a root of the cubic polynomial $z^3 - 1$. The other two roots of this polynomial are $-\frac{1}{2} - \frac{\sqrt{3}}{2}i = \bar{\omega} = \omega^2$ and $1 = \omega^0 = \omega^3$ (see Figure 1.3). The multiplication table for the set $C_3 = \{1, \omega, \omega^2\}$ under ordinary complex multiplication is

$\cdot$	1	ω	ω^2
1	1	ω	ω^2
ω	ω	ω^2	1
ω^2	ω^2	1	ω

The following is a straightforward but important fact about group multiplication, which we will need in this and some later chapters.

Proposition 1.15 Let $G = (G, *)$ be a group, and $g, h, k \in G$ be any three elements of G . Then the left and right **cancellation laws** hold:

$$g * h = g * k \implies h = k \quad (1.4)$$

$$h * g = k * g \implies h = k \quad (1.5)$$

Proof Suppose $g * h = g * k$. Multiplying both sides of this equation on the left by the inverse g^{-1} yields $g^{-1} * g * h = g^{-1} * g * k$, which gives $1 * h = 1 * k$, hence $h = k$ as required. The right cancellation law follows by a very similar argument. $\square$

Some more book-keeping: at this point we have required the existence of an identity element $e \in G$, and an inverse g^{-1} for each element $g \in G$. The following proposition confirms uniqueness of these elements.

Proposition 1.16 The identity element e of a group G is unique. That is, for any other element $f \in G$ satisfying condition G2 in Definition 1.9, we have $f = e$.

Any element g of a group G has a unique inverse g^{-1} . That is, for any other element $\bar{g}$ satisfying condition G3 in Definition 1.9, we have $\bar{g} = g^{-1}$.

Proof Suppose that $f \in G$ also satisfies the identity condition, that for any element $g \in G$, we have $f * g = g$ and $g * f = g$. In particular, $f * e = e$. But since e is also an identity, we have $f * e = f$ as well. So $f = e$.

Now suppose g^{-1} and $\bar{g}$ are two inverses for an element $g \in G$. Then $g^{-1} * g = \bar{g} * g = e$. But by condition G3 we have $g * g^{-1} = e$ as well. So

$$g^{-1} = e * g^{-1} = (\bar{g} * g) * g^{-1} = \bar{g} * (g * g^{-1}) = \bar{g} * e = \bar{g}.$$

Hence the identity element and the inverse elements are unique. $\square$

While on the subject of inverses, it's illuminating to think about what the inverse of a product of two elements looks like. Given a group G and two elements $g, h \in G$, the inverse $(g * h)^{-1} = h^{-1} * g^{-1}$. We might be tempted to assume $(g * h)^{-1} = g^{-1} * h^{-1}$ but this is not the case in general (unless G happens to be abelian, in which case we can reorder the products to our hearts' content). Remember that $(g * h)^{-1}$ has to be the unique element of G which, when multiplied by $(g * h)$, either on the left or the right, gives the identity element e . The following shows that $h^{-1} * g^{-1}$ is precisely the element we want:

$$\begin{aligned} (h^{-1} * g^{-1}) * (g * h) &= h^{-1} * (g^{-1} * g) * h = h^{-1} * h = e \\ (g * h) * (h^{-1} * g^{-1}) &= g * (h * h^{-1}) * g^{-1} = g * g^{-1} = e \end{aligned}$$

Another important observation concerning the group C_3 in Example 1.14 is that its multiplication table has essentially the same structure as the multiplication table for the cyclic group $\mathbb{Z}_3$.

+	0	1	2	·	1	ω	ω^2
0	0	1	2	1	1	ω	ω^2
1	1	2	0	ω	ω	ω^2	1
2	2	0	1	ω^2	ω^2	1	ω

In the multiplicative group of complex cube-roots of unity, the identity element is clearly 1 (since multiplying any complex number by 1 leaves it unchanged), while in $\mathbb{Z}_3$ it is 0 (since adding 0 to any integer leaves it unchanged). Similarly, ω and ω^2 behave analogously, under multiplication, to the integers 1 and 2 in modulo-3 arithmetic.

In some sense, these groups are actually the same: apart from some fairly superficial relabelling, their elements interact in the same way, and the structure of their multiplication tables are essentially identical.

More explicitly, we have a bijective correspondence

$$1 \longleftrightarrow 0 \qquad \omega \longleftrightarrow 1 \qquad \omega^2 \longleftrightarrow 2 \qquad (1.6)$$

between the elements of both groups. Actually, this structural correspondence between $\mathbb{Z}_3$ and the cube roots of unity is to be expected.

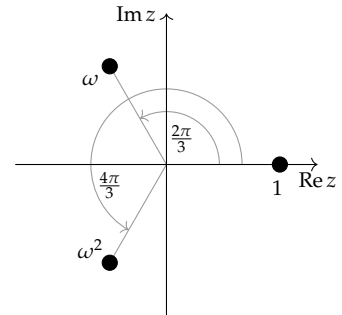


Figure 1.4: Cube roots of unity



Wikimedia Commons / Jakob Emanuel Handmann (1756)

The Swiss mathematician **Leonhard Euler** (1707–1783) (his surname is pronounced, roughly, “oiler” rather than “yuler”) was one of the most prolific mathematicians of all time. His contributions to mathematics and physics include pioneering work on analysis, number theory, graph theory, astronomy, logic, engineering and optics.

The son of a Calvinist pastor, Euler was tutored in his youth by Johann Bernoulli (1667–1748), a family friend and eminent mathematician in his own right, who encouraged him to study mathematics rather than theology at the University of Basel.

He graduated in 1726 and moved to the Imperial Academy of Sciences in St Petersburg. At this time, a suspicious political class had reasserted their control and cut scientific funding: a problem still regrettably common today.

After a near-fatal fever in 1735, Euler’s eyesight began to deteriorate, and he eventually went almost completely blind. He compensated for this with a prodigious memory (he could recite the *Aeneid* in its entirety) and his mathematical productivity was unaffected.

In 1741 he moved to Berlin at the invitation of Frederick the Great of Prussia, where he stayed for the next twenty-five years. He returned to St Petersburg in 1766, during the reign of Catherine the Great, where he lived until his death from a stroke at the age of 76.

His complete works comprise 866 known books, articles and letters. The publication of a definitive, annotated, collected edition, the *Opera Omnia*, began in 1911 and is not yet complete but has so far yielded 76 separate volumes.

Writing 1, ω and ω^2 in polar form, using **Euler’s formula**

$$e^{i\theta} = \cos \theta + i \sin \theta,$$

(see Figure 1.4) we find that

$$1 = e^0, \quad \omega = e^{\frac{2\pi i}{3}}, \quad \omega^2 = e^{\frac{4\pi i}{3}}.$$

Multiplying any two of these together and using the usual rule for products of exponents gives us a further insight into what’s going on, and enables us to write down an explicit function

$$\phi: \mathbb{Z}_3 \rightarrow C_3; \quad k \mapsto e^{\frac{2k\pi i}{3}}$$

Although the superficial appearance of a group will often give us some insight into its fundamental nature, we will be primarily interested in its underlying structure and properties. It would, therefore, be useful to have some way of saying whether two given groups are equivalent, and from the above example, the existence of a suitable bijection seems to be a good place to start. But will any bijection do? What happens if, instead of the bijection in (1.6), we use the following one?

$$1 \longleftrightarrow 1 \quad \omega \longleftrightarrow 2 \quad \omega^2 \longleftrightarrow 0 \quad (1.7)$$

The multiplication table for the relabelled group then looks like

	1	2	0
1	1	2	0
2	2	0	1
0	0	1	2

which we then rearrange to

	0	1	2
0	2	0	1
1	0	1	2
2	1	2	0

which doesn’t represent the modulo–3 addition table of the set $\{0, 1, 2\}$. So not just any bijection will do: the problem here is that the product operations don’t line up properly. However, with the bijection

$$1 \longleftrightarrow 0 \quad \omega \longleftrightarrow 2 \quad \omega^2 \longleftrightarrow 1 \quad (1.8)$$

the multiplication table for the relabelled group looks like

	0	2	1
0	0	2	1
2	2	1	0
1	1	0	2

which rearranges to

	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

which is the same as that for $\mathbb{Z}_3$.

So, we want a bijection which, like (1.8), respects the structure of the groups involved. More generally, given two groups $G = (G, *)$ and $H = (H, \circ)$ which are structurally equivalent, we want a bijection $\phi: G \rightarrow H$ such that the product in H of the images of any two elements of G is the same as the image of their product in G . This leads us to the following definition.

Definition 1.17 Two groups $G = (G, *)$ and $H = (H, \circ)$ are **isomorphic** (written $G \cong H$) if there exists a bijection (an **isomorphism**) $\phi: G \rightarrow H$ such that

$$\phi(g_1 * g_2) = \phi(g_1) \circ \phi(g_2).$$

for any $g_1, g_2 \in G$.

Later we will consider the more general case of **homomorphisms**: functions which respect group structures, but which are not necessarily bijections. For the moment, however, we are interested in isomorphisms as an explicit structural equivalence between groups.

In the group $C_3 \cong \mathbb{Z}_3$ from Example 1.14, both primitive roots $z = \omega, \omega^2$ have the property that $z^3 = 1$, but this is not true for any smaller power. That is, $n = 3$ is the smallest positive integer for which $z^n = 1$.

Definition 1.18 Let g be an element of a group G with identity element e . Then the **order** of g , denoted $|g|$, is the smallest positive integer n such that $g^n = e$. If there is no such integer n , then we say that g has **infinite order**. An element of finite order is sometimes called a **torsion element**.

Here, g^n denotes the n th power $g * \cdots * g$ of g . If, as in the case of the cyclic groups $\mathbb{Z}_n = (\mathbb{Z}_n, +)$, we are using additive notation, then we would replace g^n with ng in the above definition.

Considering the group C_3 , we remark that $|\omega| = |\omega^2| = 3$ but $|1| = 1$. In the case of $\mathbb{Z}_4$, the orders of the four elements are given by

$$|1| = |3| = 4, \quad |2| = 2, \quad |0| = 1.$$

For the order-12 cyclic group $\mathbb{Z}_{12}$, the orders are

$$\begin{aligned} |1| = |5| = |7| = |11| = 12, & \quad |2| = |10| = 6, & \quad |3| = |9| = 4, \\ |4| = |8| = 3, & \quad |6| = 2, & \quad |0| = 1. \end{aligned}$$

In all three of these examples we see that the order of the identity element is 1, and furthermore that no other element apart from the identity has order 1. This is true in general:

Proposition 1.19 *Let G be a group with identity element e . Then for any element $g \in G$ it follows that $|g| = 1$ if and only if $g = e$.*

Proof The order of e is always 1, since 1 is the smallest positive integer n for which $e^n = e$. Conversely, if $g^1 = e$ then $g = e$. $\square$

The simplest nontrivial case is that of a group where all the non-identity elements have order 2:

Proposition 1.20 *Let G be a nontrivial group, all of whose elements apart from the identity have order 2. Then G is abelian.*

Proof Let $g, h \in G$. Then by the hypothesis, $g^2 = h^2 = e$, and so $g = g^{-1}$ and $h = h^{-1}$. It follows that

$$g * h = g^{-1} * h^{-1} = (h * g)^{-1} = h * g$$

as required. $\square$

Given that an isomorphism preserves at least some aspects of the structure of a group, it is reasonable to ask how it affects the order of a given element. The answer is that it leaves it unchanged:

Proposition 1.21 *If $\phi: G \rightarrow H$ is an isomorphism, then $|\phi(g)| = |g|$ for any $g \in G$.*

Proof Suppose g has order n in G . Then $g^n = g * \cdots * g = e_G$. Therefore

$$\phi(g)^n = \phi(g) \circ \cdots \circ \phi(g) = \phi(g * \cdots * g) = \phi(g^n) = \phi(e_G) = e_H.$$

We must now check that n is the *smallest* positive integer such that $\phi(g)^n = e_H$. Suppose that there exists some $1 \leq k < n$ such that $\phi(g)^k = e_H$. Then

$$\phi(g)^k = \phi(g) \circ \cdots \circ \phi(g) = \phi(g * \cdots * g) = \phi(g^k).$$

But $g^k \neq e_G$, so $\phi(g^k) \neq \phi(e_G) = e_H$, which contradicts the assertion that $\phi(g)^k = e_H$ for some $k < n$. Therefore n is after all the smallest positive integer such that $\phi(g)^n = e_H$, and hence $|\phi(g)| = |g| = n$. $\square$

As it happens, we can construct the group $\mathbb{Z}_n$ by starting with the identity element 0 and repeatedly adding 1 to it (subject to modulo- n arithmetic). This process yields all elements of $\mathbb{Z}_n$; if we threw away all of $\mathbb{Z}_n$ except for 0, 1 and the addition operation, we could rebuild it by just adding 1 to itself enough times.

We say that the element 1 **generates** the group $\mathbb{Z}_n$, and we write $\langle 1 \rangle = \mathbb{Z}_n$. More generally:

Definition 1.22 Suppose that $g_1, g_2, \dots \in G$ are a (possibly infinite) collection of elements of some group G . Denote by $\langle g_1, g_2, \dots \rangle$ the set of elements of G which can be formed by arbitrary finite products of the elements $g_1, \dots, g_k$ and their inverses. If $G = \langle g_1, g_2, \dots \rangle$ then we say that the elements $g_1, g_2, \dots$ **generate** G , or are **generators** for G . If a group G is generated by a finite number of such elements, it is said to be **finitely generated**.

The finite cyclic groups $\mathbb{Z}_n$ can be generated by a single element, as can the (infinite) group $\mathbb{Z}$. This leads to the following definition:

Definition 1.23 A group which can be generated by a single element is said to be **cyclic**.

We will sometimes refer to $\mathbb{Z}$ as the **infinite cyclic group**. Up to isomorphism, there is only one cyclic group of a given order:

Proposition 1.24 *Suppose that G and H are (finite or infinite) cyclic groups with $|G| = |H|$. Then $G \cong H$.*

Proof Consider the infinite case first: $G = \langle g \rangle = \{g^k : k \in \mathbb{Z}\}$ and $H = \langle h \rangle = \{h^k : k \in \mathbb{Z}\}$. The elements g^j and g^k are distinct (that is, $g^j \neq g^k$) if $j \neq k$. Hence the function $\phi: G \rightarrow H$ defined by $\phi(g^k) = h^k$ is a bijection. It's also an isomorphism of groups, because

$$\phi(g^j g^k) = \phi(g^{j+k}) = h^{j+k} = h^j h^k = \phi(g^j) \phi(g^k)$$

for any $j, k \in \mathbb{Z}$. Thus $G \cong H$.

The finite case is very similar. Let $G = \langle g \rangle = \{g^k : k = 0, \dots, n-1\}$ and $H = \langle h \rangle = \{h^k : k = 0, \dots, n-1\}$, so that $|G| = |H| = n$. Then the map $\phi: G \rightarrow H$ defined by $\phi(g^k) = h^k$ is also a bijection (since as in the infinite case, the elements g^k of G are distinct for all $k = 0, \dots, n-1$). It also satisfies the condition $\phi(g^j g^k) = \phi(g^j) \phi(g^k)$ for all $0 \leq j, k \leq n-1$, and so $G \cong H$. $\square$

In particular, the finite cyclic groups in Example 1.12 are precisely the finite groups satisfying Definition 1.23. We also can derive a few general results about the order of generators in cyclic groups:

Proposition 1.25 *Let $G = \langle g \rangle$ be a cyclic group. If G is an infinite cyclic group, g has infinite order; if G is a finite cyclic group with $|G| = n$ then $|g| = n$.*

Proof If the order $|g|$ of the generator g is finite, equal to some $k \in \mathbb{N}$, say, then $g^m = g^{m+k} = g^{m+tk}$ for all integers t and m . So $G = \langle g \rangle = \{g^i : i \in \mathbb{Z}\}$ contains at most k elements. This proves the first statement, since if G is an infinite cyclic group, then g can't have finite order: if it did, G could only have a finite number of elements.

It almost proves the second statement too; all that remains is to show that a finite cyclic group G whose generator g has order k contains *exactly* k elements. This follows from the observation that $g^m = e$ if and only if m is an integer multiple of k , and hence G must contain at least k elements. Therefore G has exactly k elements. $\square$

It is also interesting to ask which elements generate all of $\mathbb{Z}_n$.

Proposition 1.26 *Let $k \in \mathbb{Z}_n$. Then k is a generator for $\mathbb{Z}_n$ if and only if $\gcd(k, n) = 1$; that is, if k and n are coprime.*

Proof From Proposition 1.25, we know that k generates all of $\mathbb{Z}_n$ if and only if it has order n . In other words, the smallest positive integer m such that $n \mid mk$ is n itself, which is the same as saying that $\gcd(k, n) = 1$, that is, k and n are coprime. $\square$

This proposition tells us that any integer $k \in \{0, \dots, n-1\}$ that is coprime to n can generate the entirety of the finite cyclic group $\mathbb{Z}_n$. The number of possible generators of $\mathbb{Z}_n$ is sometimes denoted $\phi(n)$; this is **Euler's totient function**.

Given two sets X and Y , we can form their **cartesian product** $X \times Y$, which we define to be the set

$$X \times Y = \{(x, y) : x \in X, y \in Y\}$$

of **ordered pairs** of elements of X and Y . Since a group is, fundamentally, a set with some additional structure defined on it, can we use this cartesian product operation to make new groups by combining two or more smaller ones in a similar way? The answer is yes:⁵

⁵ There are various other group structures that can be defined on the set $G \times H$. Two particularly important examples are the **semidirect product** $H \rtimes G$ and the **wreath product** $H \wr G$, both of which we will meet in Section 7.4.

Definition 1.27 Given two groups $G = (G, *)$ and $H = (H, \circ)$, their **direct product** is the group $G \times H = (G \times H, \bullet)$ whose underlying set is the cartesian product $G \times H$ of the sets G and H , and group operation $\bullet$ given by

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

If the groups G and H are written in additive notation (especially if they are abelian) then we call the corresponding group the **direct sum** and denote it $G \oplus H$.

Before proceeding any further, we must first check that this operation does in fact yield a group, rather than just a set with some insufficiently group-like structure defined on it. Fortunately, it does:

Proposition 1.28 *Given two groups $G = (G, *)$ and $H = (H, \circ)$, their direct product $G \times H$ is also a group, and is abelian if and only if G and H both are.*

Proof This is fairly straightforward, and really just requires us to check the group axioms. We begin with associativity:

$$\begin{aligned} ((g_1, h_1) \bullet (g_2, h_2)) \bullet (g_3, h_3) &= (g_1 * g_2, h_1 \circ h_2) \bullet (g_3, h_3) \\ &= ((g_1 * g_2) * g_3, (h_1 \circ h_2) \circ h_3) \\ &= (g_1 * (g_2 * g_3), h_1 \circ (h_2 \circ h_3)) \\ &= (g_1, h_1) \bullet (g_2 * g_3, h_2 \circ h_3) \\ &= (g_1, h_1) \bullet ((g_2, h_2) \bullet (g_3, h_3)) \end{aligned}$$

for all $g_1, g_2, g_3 \in G$ and $h_1, h_2, h_3 \in H$. Next we need to verify the existence of an identity element. If e_G is the identity element in G and e_H is the identity element in H , then $e_{G \times H} = (e_G, e_H)$ is the identity element in $G \times H$, since $(e_G, e_H) \bullet (g, h) = (e_G * g, e_H \circ h) = (g, h)$ and $(g, h) \bullet (e_G, e_H) = (g * e_G, h \circ e_H) = (g, h)$ for any $g \in G$ and $h \in H$. We define the inverses in $G \times H$ in a similar way, with $(g, h)^{-1} = (g^{-1}, h^{-1})$, since $(g^{-1}, h^{-1}) \bullet (g, h) = (g^{-1} * g, h^{-1} \circ h) = (e_G, e_H)$ and

$(g, h) \bullet (g^{-1}, h^{-1}) = (g * g^{-1}, h \circ h^{-1}) = (e_G, e_H)$. This completes the proof that $G \times H$ is a group. To prove the second part, we observe that

$$(g_1, h_1) \bullet (g_2, h_2) = (g_1 * g_2, h_1 \circ h_2)$$

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

These expressions are equal (and hence $G \times H$ is abelian) if and only if both G and H are abelian. $\square$

Example 1.29 The group $\mathbb{Z}_2 \oplus \mathbb{Z}_3$ has the multiplication table

	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)
(0,0)	(0,0)	(0,1)	(0,2)	(1,0)	(1,1)	(1,2)
(0,1)	(0,1)	(0,2)	(0,0)	(1,1)	(1,2)	(1,0)
(0,2)	(0,2)	(0,0)	(0,1)	(1,2)	(1,0)	(1,1)
(1,0)	(1,0)	(1,1)	(1,2)	(0,0)	(0,1)	(0,2)
(1,1)	(1,1)	(1,2)	(1,0)	(0,1)	(0,2)	(0,0)
(1,2)	(1,2)	(1,0)	(1,1)	(0,2)	(0,0)	(0,1)

This group is isomorphic to the cyclic group $\mathbb{Z}_6$, by the isomorphism

$$\begin{array}{lll} (0,0) \mapsto 0 & (1,1) \mapsto 1 & (0,1) \mapsto 2 \\ (1,0) \mapsto 3 & (0,2) \mapsto 4 & (1,2) \mapsto 5 \end{array}$$

Alternatively, note that $\mathbb{Z}_2 \oplus \mathbb{Z}_3 = \langle (1,1) \rangle$ and so must be cyclic by Definition 1.23, and since it has six elements, it must be isomorphic to $\mathbb{Z}_6$ by Proposition 1.24.

Example 1.30 The multiplication table for the group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ is

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	(0,1)	(0,0)	(1,1)	(1,0)
(1,0)	(1,0)	(1,1)	(0,0)	(0,1)
(1,1)	(1,1)	(1,0)	(0,1)	(0,0)

This group is not cyclic: every nontrivial element has order 2, so no single element can generate the entire group. Hence $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \not\cong \mathbb{Z}_4$.

The group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ is named the **Klein 4-group** or **Viergruppe** after the German mathematician Felix Klein (1849–1925). We will see in a little while that it describes the symmetry of a rectangle.

Example 1.31 The group $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$ has eight elements, as do the groups $\mathbb{Z}_8$ and $\mathbb{Z}_4 \oplus \mathbb{Z}_2$. If we write out the multiplication tables, though, we find that there are important structural differences between them. Neither $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$ nor $\mathbb{Z}_4 \oplus \mathbb{Z}_2$ are cyclic, since no single element generates the entire group. Nor are they isomorphic to each other, since $\mathbb{Z}_4 \oplus \mathbb{Z}_2 = \langle (0,1), (1,0) \rangle$ but no two elements of $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$ generate the entire group.



Oberwolfach Photo Collection

The German mathematician **Felix Klein** (1849–1925) began his research career under the supervision of the physicist and geometer Julius Plücker (1801–1868), receiving his doctorate from the University of Bonn in 1868, and completing Plücker's treatise on line geometry after the latter's death. Appointed professor at the University of Erlangen in 1872 at the age of 23, Klein instigated his *Erlangen Programme* to classify geometries via their underlying symmetry groups. In 1886 he was appointed to a chair at Göttingen, where he remained until his retirement in 1913, doing much to re-establish it as the world's leading centre for research in mathematics.

He also did much to further academic prospects for women. In 1895 his student Grace Chisholm Young (1868–1944) became the first woman to receive a doctorate by thesis from any German university (the Russian mathematician Sofya Kovalevskaya (1850–1891) was awarded hers for published work in 1874). Klein was also instrumental, along with David Hilbert (1862–1943), in securing an academic post for the algebraist Emmy Noether (1882–1935).

The **Klein bottle**, a non-orientable closed surface which cannot be embedded in 3-dimensional space, bears his name.

Alternatively, there is no isomorphism $\phi: \mathbb{Z}_4 \oplus \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, since by Proposition 1.21, any such isomorphism must preserve the order of each element; however the element $(1, 0) \in \mathbb{Z}_4 \oplus \mathbb{Z}_2$ has order 4, but there is no order-4 element in $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$ for it to map to, since every nontrivial element in the latter group has order 2.

A similar argument shows that $\mathbb{Z}_8$ (which has four order-8 elements) cannot be isomorphic to either of the other two groups.

Why is $\mathbb{Z}_2 \oplus \mathbb{Z}_3 \cong \mathbb{Z}_6$ but $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \not\cong \mathbb{Z}_4$, and none of the order-8 groups in the above example are isomorphic to each other? The answer is given by the following proposition.

Proposition 1.32 $\mathbb{Z}_m \oplus \mathbb{Z}_n \cong \mathbb{Z}_{mn}$ if and only if $\gcd(m, n) = 1$; that is, m and n are coprime.

Proof The key is to examine the subset $\langle (1, 1) \rangle$ of $\mathbb{Z}_m \oplus \mathbb{Z}_n$ generated by the element $(1, 1)$. This is the set of all ordered pairs from $\mathbb{Z}_m \oplus \mathbb{Z}_n$ that we get by adding together a finite number of copies of $(1, 1)$. Since both m and n are finite, the first coordinate will cycle back round to 0 after m additions, while the second will do so after n additions.

Suppose that m and n are coprime. Then after m additions the first coordinate will be 0 but the second won't. Similarly, after n additions the second coordinate will be 0 but the first won't. To get back to $(0, 0)$ we need to add $(1, 1)$ to itself a number of times that contains both m and n as factors; the smallest such number is the lowest common multiple $\text{lcm}(m, n)$. Since m and n are coprime, $\text{lcm}(m, n) = mn$, so this subset $\langle (1, 1) \rangle$ has mn distinct elements. But $\mathbb{Z}_m \oplus \mathbb{Z}_n$ also has mn elements, thus $(1, 1)$ generates the entirety of $\mathbb{Z}_m \oplus \mathbb{Z}_n$. So $\mathbb{Z}_m \oplus \mathbb{Z}_n$ is cyclic and is hence isomorphic to $\mathbb{Z}_{mn}$ by Proposition 1.24.

To prove the converse, suppose that $\gcd(m, n) = d > 1$. Then both m and n are divisible by d , and so $\text{lcm}(m, n) = \frac{mn}{d}$. This means that the subset $\langle (1, 1) \rangle$ has $\frac{mn}{d}$ elements and hence can't be isomorphic to $\mathbb{Z}_{mn}$ but instead is isomorphic to $\mathbb{Z}_{mn/d}$.

In fact, no element $(a, b) \in \mathbb{Z}_m \oplus \mathbb{Z}_n$ can generate the whole of the group, since

$$(a, b) + \cdots + (a, b) = \left(\frac{mn}{d}a, \frac{mn}{d}b\right) = \left(\frac{n}{d}ma, \frac{m}{d}nb\right) = (0, 0).$$

So $\mathbb{Z}_m \oplus \mathbb{Z}_n$ isn't cyclic and can't be isomorphic to $\mathbb{Z}_{mn}$. □

In Section 5.3 we'll use this result as part of the classification theorem for finitely-generated abelian groups,^{6,7} and in Section 9.1 we will use it to help prove the Chinese Remainder Theorem for rings.⁸

⁶ Theorem 5.48, page 169.

⁷ Theorem 5.63, page 178.

⁸ Proposition 9.8, page 348.

1.2 Matrices

ALL OF THE CONCRETE EXAMPLES of groups we've met so far have been numerical in nature, and also abelian. In this section we will see some examples of nonabelian groups formed from matrices.

Multiplication of $n \times n$ square matrices is well-defined and associative, and the $n \times n$ identity matrix

$$I_n = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

satisfies the identity condition. Not every square matrix has an inverse, but as long as we restrict our attention to those that do, we should be able to find some interesting examples of groups.

Example 1.33 The matrices

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}, \quad B = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}.$$

forms a group with multiplication table

$\cdot$	I	A	B
I	I	A	B
A	A	B	I
B	B	I	A

which is clearly isomorphic to the table for the group $\mathbb{Z}_3$ that we met a little while ago, under the isomorphism

$$I \mapsto 0, \quad A \mapsto 1, \quad B \mapsto 2.$$

We now use the fact that matrix multiplication is, in general, noncommutative to give us our first examples of nonabelian groups.

Example 1.34 Let $GL_n(\mathbb{R})$ denote the set of invertible, $n \times n$ matrices with real entries. This set forms a group (the **general linear group**) under the usual matrix multiplication operation.

The set $GL_n(\mathbb{R})$ is closed under multiplication (since any two invertible matrices A and B form a product AB with inverse $B^{-1}A^{-1}$), matrix multiplication is associative, the $n \times n$ identity matrix $I_n \in GL_n(\mathbb{R})$, and each matrix $A \in GL_n(\mathbb{R})$ has an inverse $A^{-1} \in GL_n(\mathbb{R})$. We can specialise this example to give three other (also nonabelian) groups which are contained within $GL_n(\mathbb{R})$:

Example 1.35 The **special linear group** $SL_n(\mathbb{R})$ is the group of $n \times n$ real matrices with determinant 1.

If only I knew how to get mathematicians interested in transformation groups and their applications to differential equations. I am certain, absolutely certain, that these theories will some time in the future be recognised as fundamental. When I wish such a recognition sooner, it is partly because then I could accomplish ten times more.

— Marius Sophus Lie (1842–1899),
letter written in 1884 to Adolf Mayer
(1843–1942)

This matrix group $\{I, A, B\}$ is an example of quite a powerful and important concept called a **representation**: broadly speaking, a formulation of a given group as a collection of matrices which behave in the same way. We will study this idea in a little more detail in Section 8.A.



Wikimedia Commons / Ludwik Szacinski (1844–1894)

After graduating in 1865 from the University of Christiania (now Oslo) with a general science degree, **Marius Sophus Lie** (1842–1899) dabbled in various subjects, including astronomy, zoology and botany, before settling on mathematics. In 1869, on the strength of his first published paper, *Repräsentation der Imaginären der Plangeometrie*, he won a scholarship to travel to Berlin and Paris, where he met and worked with several important mathematicians including Felix Klein (1849–1925) and Camille Jordan (1838–1922).

After the outbreak of the Franco-Prussian War in July 1870, he left for Italy but was arrested in Fontainebleau on suspicion of being a German spy and was only released due to the intervention of the French mathematician Jean Gaston Darboux (1842–1917).

Returning to Norway, he was awarded a doctorate by the University of Christiania in July 1872, for a thesis entitled *On a class of geometric transformations*, and subsequently appointed to a chair.

Over the next few decades he made many important contributions to geometry and algebra, many in collaboration with Klein and Friedrich Engel (1861–1941). In 1886 he moved to Leipzig, succeeding Klein (who had just been appointed to a chair at Göttingen) but suffered a nervous breakdown in late 1889. This, together with his deteriorating physical health and the acrimonious disintegration of his professional relationships with Klein and Engel overshadowed the rest of his career. He returned to Christiania in 1898, and died from pernicious anaemia in 1899.

Example 1.36 Let $O_n(\mathbb{R})$ denote the set of $n \times n$ **orthogonal** real matrices. That is, invertible matrices A whose inverse is equal to their transpose, $A^{-1} = A^T$. This set forms a group (the **orthogonal group**) under matrix multiplication.

Example 1.37 The **special orthogonal group** $SO_n(\mathbb{R})$ is the group of $n \times n$ orthogonal real matrices with determinant 1.

There's nothing in particular restricting us to real matrices either. The following two examples, of matrices with complex entries, have particular relevance to quantum mechanics.

Example 1.38 Let U_n denote the set of $n \times n$ **unitary** matrices. That is, invertible matrices A whose inverse is equal to their **conjugate transpose** $A^\dagger$, the matrix formed by taking the transpose A^T and replacing every entry a_{ij} with its complex conjugate $\bar{a}_{ij}$. This set U_n forms a group (the **unitary group**) under matrix multiplication.

The unitary group U_1 is isomorphic to the **circle group**: the multiplicative group of complex numbers z with unit modulus $|z| = 1$.

Example 1.39 Let SU_n denote the group of $n \times n$ unitary matrices with determinant 1. This is the **special unitary group**.

In the Standard Model of particle physics, U_1 describes the quantum behaviour of the electromagnetic force, SU_2 the weak nuclear force, and SU_3 the strong nuclear force.

Example 1.40 A **symplectic matrix** is a $2n \times 2n$ square matrix M satisfying the condition

$$M^T J_{2n} M = J_{2n},$$

where J_{2n} is the $2n \times 2n$ block matrix

$$J_{2n} = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix},$$

with I_n denoting the $n \times n$ identity matrix. So, for example,

$$J_4 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}.$$

Equivalently, a $2n \times 2n$ block matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is **symplectic** if the $n \times n$ submatrices A, B, C and D satisfy the conditions

$$\begin{aligned} A^T C &= C^T A, & A^T D - C^T B &= I, \\ B^T D &= D^T B, & D^T A - B^T C &= I. \end{aligned}$$

The **symplectic group** $Sp_{2n}(\mathbb{R})$ is composed of all $2n \times 2n$ symplectic matrices with real entries.

These groups $GL_n(\mathbb{R})$, $SL_n(\mathbb{R})$, $O_n(\mathbb{R})$, $SO_n(\mathbb{R})$, $Sp_{2n}(\mathbb{R})$, U_n and SU_n are, for $n > 1$ at least, all infinite and nonabelian. But they also have a continuous structure inherited from $\mathbb{R}$, which means that we can regard them as being locally similar to $\mathbb{R}^m$ for some value of m . Such an object is called a **manifold**, and a group which has a compatible manifold structure in this way is known as a **Lie group**, after the Norwegian mathematician Marius Sophus Lie (1842–1899).

Example 1.41 Let I , A and B be as in Example 1.33, and let

$$C = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad D = \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}, \quad E = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & \frac{1}{2} \end{bmatrix}.$$

Then the set $\{I, A, B, C, D, E\}$ forms a group with multiplication table

$\cdot$	I	A	B	C	D	E
I	I	A	B	C	D	E
A	A	B	I	D	E	C
B	B	I	A	E	C	D
C	C	E	D	I	B	A
D	D	C	E	A	I	B
E	E	D	C	B	A	I

This group can be seen to be nonabelian: its multiplication table is asymmetric in the leading (top-left to bottom-right) diagonal.

Example 1.42 The three unitary matrices

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

are known as **Pauli's matrices**, and are particularly relevant in particle physics (where they represent observables relating to the spin of spin- $\frac{1}{2}$ particles such as protons, neutrons and electrons) and quantum computing (where they represent an important class of single-qubit operations). Now consider the matrices

$$I = i\sigma_z = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \quad J = i\sigma_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad K = i\sigma_x = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$$

together with the 2×2 identity matrix, which in this instance we call E . Let $-E$, $-I$, $-J$ and $-K$ be the negative scalar multiples of these matrices. The group formed by these eight elements is called the **quaternion group** Q_8 ; its multiplication table is shown in Table 1.2.

The quaternion group Q_8 is not isomorphic to any of the three other order-8 groups we've met so far, since they were abelian and this isn't. Later, we will study some structures related to this group, such as the quaternion ring $\mathbb{H}$ and the Hurwitz integers.⁹



© 1933 CERN / Wolfgang Pauli Archive

The Austrian theoretical physicist **Wolfgang Pauli** (1900–1958) was one of the pioneers of quantum mechanics and particle physics. His contributions include the **Pauli Exclusion Principle**, which states that no two electrons (more generally, any two fermions) can exist in the same quantum state at the same time, for which he was awarded the Nobel Prize for Physics in 1945. In 1930, while working on the problem of beta decay, he postulated the existence of a new particle (the **neutrino**) whose existence was confirmed experimentally 26 years later.

A notorious perfectionist, he would often dismiss work he considered substandard as “falsch” (“wrong”) or “ganz falsch” (“completely wrong”); famously he once remarked of one research paper “es ist nicht einmal falsch” (“it is not even wrong”).

	E	I	J	K	$-E$	$-I$	$-J$	$-K$
E	E	I	J	K	$-E$	$-I$	$-J$	$-K$
I	I	E	K	$-J$	$-I$	E	$-K$	J
J	J	$-K$	E	I	$-J$	K	E	$-I$
K	K	J	$-I$	E	$-K$	$-J$	I	E
$-E$	$-E$	$-I$	$-J$	$-K$	E	I	J	K
$-I$	$-I$	E	$-K$	J	I	E	K	$-J$
$-J$	$-J$	K	E	$-I$	J	$-K$	E	I
$-K$	$-K$	$-J$	I	E	K	J	$-I$	$-E$

Table 1.2: Multiplication table for the quaternion group Q_8

⁹ Example 8.12, page 322.

Symmetry: That which we see at a glance; based on the fact that there is no reason to do otherwise; and based also on the human form, from where it follows that symmetry is necessary only in width, not in height or depth.

— Blaise Pascal (1623–1662),
Pensées (1669) 1:28

$\cdot$	I	A	B
I	I	A	B
A	A	B	I
B	B	I	A

Table 1.3: The multiplication table for the group in Example 1.33

$\cdot$	I	A	B	C	D	E
I	I	A	B	C	D	E
A	A	B	I	D	E	C
B	B	I	A	E	C	D
C	C	E	D	I	B	A
D	D	C	E	A	I	B
E	E	D	C	B	A	I

Table 1.4: The multiplication table for the group in Example 1.41

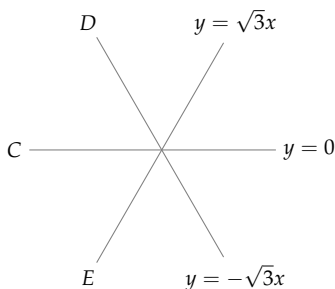


Figure 1.5: Reflections represented by the matrices C , D and E

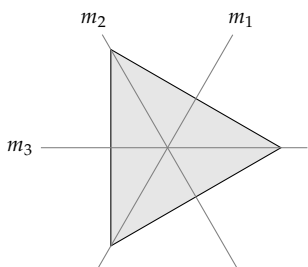


Figure 1.6: Axes of symmetry of an equilateral triangle

1.3 Symmetries

We know from basic linear algebra that matrices represent linear transformations defined on some vector space V . The matrices

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}, \quad B = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{3}}{2} & -\frac{1}{2} \end{bmatrix}$$

from Example 1.33 are 2×2 real matrices, so we can regard them as representing linear maps $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$. That is, transformations in the plane that fix the origin and satisfy the linearity conditions

$$f(\mathbf{v} + \mathbf{w}) = f(\mathbf{v}) + f(\mathbf{w}) \quad \text{and} \quad f(k\mathbf{v}) = kf(\mathbf{v})$$

for any vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$ and any real scalar $k \in \mathbb{R}$.

So what do these three 2×2 matrices do to points (or position vectors) in the plane $\mathbb{R}^2$? The identity matrix I leaves everything unchanged, while the matrices A and B represent rotations through, respectively, angles of $\frac{2\pi}{3}$ and $\frac{4\pi}{3}$ (or $-\frac{2\pi}{3}$) about the origin. The multiplication table in Table 1.3 tells us how these rotations interact with each other, and completely describes the geometry of the system. As noted earlier, this structure is equivalent to the operation of addition in modulo-3 arithmetic.

Now look at the six-element, nonabelian matrix group from Example 1.41 (see Table 1.4). These, again, are 2×2 real matrices, and can therefore be viewed as representing linear transformations in $\mathbb{R}^2$. As before, we have the identity matrix I and the two rotation matrices A and B , but we also have three further matrices. A bit of examination or experimentation shows that these matrices represent reflections in lines passing through the origin: C is reflection in the x -axis, D is reflection in the line $y = -\sqrt{3}x$, and E is reflection in the line $y = \sqrt{3}x$. This matrix group tells us how six specific geometric transformations interact with each other. For example, the identity $CE = A$ tells us that reflecting first in the x -axis and then in the line $y = \sqrt{3}x$ is the same as an anticlockwise rotation through an angle of $\frac{2\pi}{3}$.

This geometric approach is a useful and illuminating way of thinking about groups. To take this geometric viewpoint a little further, observe that these six transformations are exactly those which correspond to the symmetries of an equilateral triangle: order-3 rotational symmetry and three axes of reflection symmetry (see Figure 1.6). That is, this group completely describes the symmetry of an equilateral triangle.

This group is the **dihedral group**, and there are two common conventions for what to call it. Some mathematicians call it D_6 , because it comprises six elements, while others call it D_3 because it describes the symmetry of a regular 3-sided polygon. In this book, we'll adopt the

latter convention, but you should be aware that some books follow the other one. Its multiplication table is shown in Table 1.5.

More generally, we can do the same thing with a regular n -gon. This has order- n rotational symmetry, and n axes of symmetry. In the case where n is odd, each of these axes passes through the midpoint of one side and the opposite vertex. If n is even, then half of these axes pass through pairs of opposite vertices, and the other half pass through the midpoints of opposite sides.

Definition 1.43 The **dihedral group**

$$D_n = \{e, r, \dots, r^{n-1}, m_1, \dots, m_n\}$$

is the symmetry group of the regular plane n -gon. Here, e is the identity, r represents an anticlockwise rotation through an angle of $\frac{2\pi}{n}$, and m_k denotes a reflection in the line which makes an angle of $\frac{k\pi}{n}$ with the horizontal.

Here's another example: the dihedral group D_4 , which describes the symmetry of a square.

Example 1.44 The dihedral group $D_4 = \{e, r, r^2, r^3, m_1, m_2, m_3, m_4\}$ consists of the eight possible symmetry transformations on the square. Here, e denotes the identity transformation, r is an anticlockwise rotation through an angle of $\frac{\pi}{2}$, r^2 is therefore a rotation through an angle of π and r^3 a clockwise rotation through an angle of $\frac{\pi}{2}$, m_1 and m_3 are reflections in the square's diagonals, and m_2 and m_4 are reflections in, respectively, the vertical and horizontal axes. The multiplication table is shown in Table 1.6.

A rectangle has fewer symmetries than a square, and as we might expect, its symmetry group is simpler:

Example 1.45 Let $V_4 = \{e, r, h, v\}$ be the symmetry group of a (non-square) rectangle, where e is as usual the identity, r denotes a rotation through an angle of π about the centre point, and h and v are, respectively, reflections in the horizontal and vertical axes. Then we obtain the group with multiplication table shown in Table 1.7. This group is isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ from Example 1.30, via the isomorphism

$$e \longleftrightarrow (0,0) \quad r \longleftrightarrow (1,1) \quad h \longleftrightarrow (1,0) \quad v \longleftrightarrow (0,1)$$

and is hence the Klein 4-group in a very superficial disguise.

We can also think of this group as the dihedral group D_2 , the symmetry group of a 2-sided polygon or **bigon**, the lens-shaped figure depicted in Figure 1.9.

The dihedral groups D_n , the Klein 4-group V_4 and the group in

$\cdot$	e	r	r^2	m_1	m_2	m_3
e	e	r	r^2	m_1	m_2	m_3
r	r	r^2	e	m_3	m_1	m_2
r^2	r^2	e	r	m_2	m_3	m_1
m_1	m_1	m_2	m_3	e	r	r^2
m_2	m_2	m_3	m_1	r^2	e	r
m_3	m_3	m_1	m_2	r	r^2	e

Table 1.5: The multiplication table for the dihedral group D_3

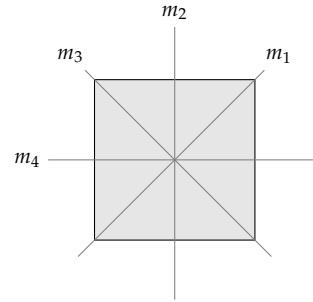


Figure 1.7: Axes of symmetry of the square

$\cdot$	e	r	r^2	r^3	m_1	m_2	m_3	m_4
e	e	r	r^2	r^3	m_1	m_2	m_3	m_4
r	r	r^2	r^3	e	m_4	m_1	m_2	m_3
r^2	r^2	r^3	e	r	m_3	m_4	m_1	m_2
r^3	r^3	e	r	r^2	m_2	m_3	m_4	m_1
m_1	m_1	m_2	m_3	m_4	e	r	r^2	r^3
m_2	m_2	m_3	m_4	m_1	r^3	e	r	r^2
m_3	m_3	m_4	m_1	m_2	r^2	r^3	e	r
m_4	m_4	m_1	m_2	m_3	r	r^2	r^3	e

Table 1.6: The multiplication table for the dihedral group D_4

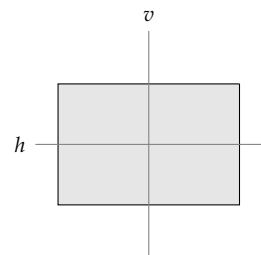


Figure 1.8: Axes of symmetry of a rectangle

$\cdot$	e	r	h	v
e	e	r	h	v
r	r	e	v	h
h	h	v	e	r
v	v	h	r	e

Table 1.7: Multiplication table for symmetry group of a rectangle

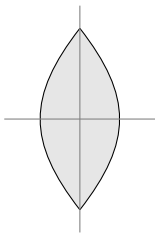
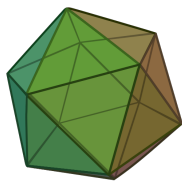
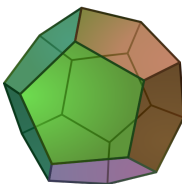
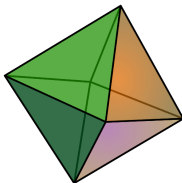
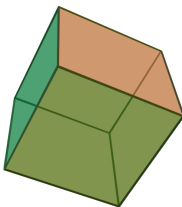
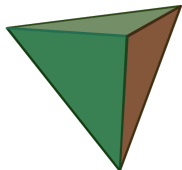


Figure 1.9: Symmetries of a bigon

Figure 1.10: The five regular polyhedra
Wikimedia Commons / User:Cyp

Example 1.33, as well as the orthogonal groups $O_n(\mathbb{R})$ and $SO_n(\mathbb{R})$ are examples of **isometry groups**: groups of transformations which preserve lengths and distances.

Definition 1.46 Given some set $S \subseteq \mathbb{R}^n$, define $\text{Isom}^+(S)$ to be the group of all **direct isometries** of S : length-preserving transformations which also preserve the orientation of S and the underlying space $\mathbb{R}^n$. Define $\text{Isom}(S)$ to be the full **isometry group** of S , consisting of the direct isometries together with the **opposite isometries**: those which reverse the orientation of S and $\mathbb{R}^n$.

In this terminology, if we let P_n denote the regular n -sided polygon, then $\text{Isom}^+(P_n) \cong \mathbb{Z}_n$ and $\text{Isom}(P_n) = D_n$.

Definition 1.47 Let $E_2^+ = \text{Isom}^+(\mathbb{R}^2)$ be the group of all orientation-preserving isometries of the plane $\mathbb{R}^2$. Then E_2^+ consists of translations and rotations. The full isometry group $E_2 = \text{Isom}(\mathbb{R}^2)$ also includes reflections and **glide reflections** (a reflection in some line, followed by a translation parallel to the same line). This group E_2 is called the (two-dimensional) **Euclidean group**.

If we let P_3 be the subset of $\mathbb{R}^2$ consisting of an equilateral triangle centred on the origin, then the dihedral group D_3 consists of precisely those six elements of $E_2 = \text{Isom}(\mathbb{R}^2)$ which map P_3 to itself, and the group in Example 1.33 consists of exactly those three direct isometries in $E_2^+ = \text{Isom}^+(\mathbb{R}^2)$ which map P_3 to itself. We will return to this viewpoint in the next chapter when we look at subgroups.

Returning to the isometry groups of Euclidean space $\mathbb{R}^n$, we find that things become progressively more complicated in higher dimensions:

Definition 1.48 Let $E_3^+ = \text{Isom}^+(\mathbb{R}^3)$ be the group of orientation-preserving isometries of $\mathbb{R}^3$. This group consists of the identity, translations, rotations about some axis, and **screw operations** (rotations about some axis followed by a translation along that axis). The full isometry group $E_3 = \text{Isom}(\mathbb{R}^3)$ also includes reflections in some plane, **glide reflections** (reflections in a plane followed by a translation parallel to that plane), **roto-reflections** (rotation about an axis followed by reflection in a plane perpendicular to that axis) and **inversion** in a point (in which every point in $\mathbb{R}^3$ is mapped to an equidistant point on the other side of the chosen fixed point). This group E_3 is the three-dimensional **Euclidean group**.

Just as we obtained interesting examples of groups by restricting ourselves to elements of E_2 or E_2^+ which mapped some subset of $\mathbb{R}^2$ to itself, it turns out that we can obtain further interesting examples by doing something similar with the groups E_3 and E_3^+ .

There are five regular polyhedra in three dimensions: the tetrahedron, the cube, the octahedron, the dodecahedron and the icosahedron. These have, respectively, four, six, eight, twelve and twenty faces, and are depicted in Figure 1.10. We will study their symmetry groups, and those of their higher-dimensional analogues, in more detail in Section 5.C but for the moment we'll look at the simplest case: the direct isometry group of the tetrahedron.

Example 1.49 The tetrahedron Δ^3 has twelve direct isometries:

- The identity isometry e
- Eight rotations $r_i^\pm$ through angles of $\pm \frac{2\pi}{3}$ around one of four axes ($i = 1, \dots, 4$) passing through a vertex and the centre of the opposite face
- Three rotations s_A, s_B and s_C through an angle of π around one of three axes passing through the midpoints of opposite edges

The multiplication table for this group $\text{Isom}^+(\Delta^3)$ is as follows:

	e	r_1^+	r_1^-	r_2^+	r_2^-	r_3^+	r_3^-	r_4^+	r_4^-	s_A	s_B	s_C
e	e	r_1^+	r_1^-	r_2^+	r_2^-	r_3^+	r_3^-	r_4^+	r_4^-	s_A	s_B	s_C
r_1^+	r_1^+	r_1^-	e	r_4^-	s_B	r_2^-	s_A	r_3^-	s_C	r_4^+	r_3^+	r_2^+
r_1^-	r_1^-	e	r_1^+	s_C	r_3^+	s_B	r_4^+	s_A	r_2^+	r_3^-	r_2^-	r_4^-
r_2^+	r_2^+	r_3^-	s_B	r_2^-	e	r_4^-	s_C	r_1^-	s_A	r_3^+	r_4^+	r_1^+
r_2^-	r_2^-	s_C	r_4^+	e	r_2^+	s_A	r_1^+	s_B	r_3^+	r_4^-	r_1^-	r_3^-
r_3^+	r_3^+	r_4^-	s_A	r_1^-	s_C	r_3^-	e	r_2^-	s_B	r_2^+	r_1^+	r_4^+
r_3^-	r_3^-	s_B	r_2^+	s_A	r_4^+	e	r_3^+	s_C	r_1^+	r_1^-	r_4^-	r_2^-
r_4^+	r_4^+	r_2^-	s_C	r_3^-	s_A	r_1^-	s_B	r_4^-	e	r_1^+	r_2^+	r_3^+
r_4^-	r_4^-	s_A	r_3^+	s_B	r_1^+	s_C	r_2^+	e	r_4^+	r_2^-	r_3^-	r_1^-
s_A	s_A	r_3^+	r_4^-	r_4^+	r_3^-	r_1^+	r_2^-	r_2^+	r_1^-	e	s_C	s_B
s_B	s_B	r_2^+	r_3^-	r_1^+	r_4^-	r_4^+	r_1^-	r_3^+	r_2^-	s_C	e	s_A
s_C	s_C	r_4^+	r_2^-	r_3^+	r_1^-	r_2^+	r_4^-	r_1^+	r_3^-	s_B	s_A	e

In this table, the operation in row a and column b is the product ab ; that is, the isometry a followed by the isometry b .

The full isometry group $\text{Isom}(\Delta^3)$ also includes twelve opposite isometries: six order-2 reflections $m_{12}, m_{13}, m_{14}, m_{23}, m_{24}$ and m_{34} (in planes passing along one edge and through the midpoint of the opposite edge) and six other roto-reflection isometries $t_A^\pm, t_B^\pm$ and $t_C^\pm$ formed by rotating the tetrahedron through an angle of $\pm \frac{\pi}{2}$ about one of the three axes s_A, s_B or s_C and then reflecting in a plane perpendicular to that axis. (The type- t isometries may also be generated by pairs of type- m reflections.)

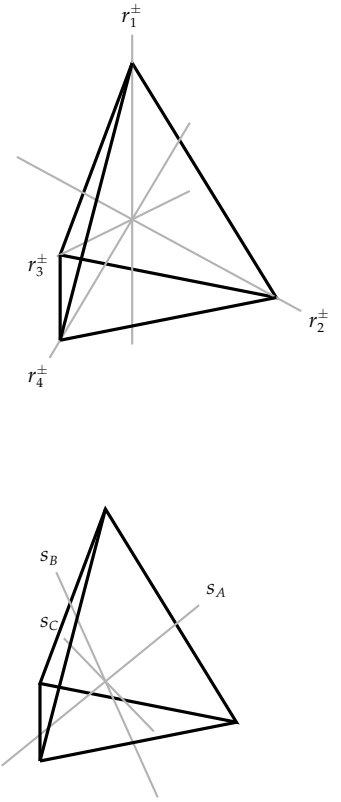


Figure 1.11: Direct isometries of the tetrahedron Δ^3 : axes for order-3 rotations (top) and order-2 rotations (bottom)

Poetical scene with surprisingly chaste
Lord Archer vegetating (3,3,8,12)

— Araucaria (John Graham)
(1921–2013),
Cryptic Crossword 22 103, *The
Guardian*, 11 January 2001

Stifle is an anagram of itself.

— Anonymous

1.4 Permutations

LOOK AGAIN AT THE ROTATION GROUP $R_3 = \text{Isom}^+(\Delta^2)$ of the equilateral triangle, from Example 1.33. Fundamentally, all these transformations really do is swap the vertices round (or, in the case of the identity transformation, not). What happens to the rest of the triangle is determined completely by where the vertices go. So we could describe the elements of R_3 , the rotation transformations on the equilateral triangle, solely as permutations of the set $\{0, 1, 2\}$.

Similarly, the dihedral group D_3 can also be regarded as permutations of the set $\{0, 1, 2\}$. The identity transformation leaves all three points unchanged, the clockwise and anticlockwise rotations permute them cyclically, and each of the three reflections leave one point unchanged and swap the other two points round.

Definition 1.50 A permutation σ of a set X is a bijection $\sigma: X \rightarrow X$.

Given a set X , we can form the set $\text{Sym}(X)$ of all possible permutations of X . In order to turn this into a group, we need to impose a suitable binary operation on it, and the one we choose is that of composition. Viewing two permutations $\sigma, \tau \in \text{Sym}(X)$ as bijections $\sigma, \tau: X \rightarrow X$, their composites $\sigma \circ \tau: X \rightarrow X$ and $\tau \circ \sigma: X \rightarrow X$ are certainly well-defined. These composites, happily, are also permutations in their own right, since the composite of two bijections is itself a bijection. If we view σ and τ as (possibly different) ways of shuffling or reordering the elements of X , then if we do σ and then do τ to the result, the combined effect is also a (possibly different) way of shuffling the elements of X .

So, we have a binary operation defined on $\text{Sym}(X)$. This operation is associative, since composition of functions is associative. There is an identity element: the identity permutation $\iota: X \rightarrow X$ which maps every element to itself. Also, every permutation σ has an inverse σ^{-1} , which can be regarded as either the inverse function $\sigma^{-1}: X \rightarrow X$ or as the permutation which puts all of the elements of X back to how they were before we applied σ . Hence $\text{Sym}(X)$ forms a group:

Definition 1.51 Let X be a (possibly infinite) set. The group $\text{Sym}(X)$ of all permutations $\sigma: X \rightarrow X$ is called the **symmetric group** on X . If $X = \{1, \dots, n\}$ is a finite set consisting of n elements, then we call $\text{Sym}(X)$ the **symmetric group** on n objects and denote it S_n .

Proposition 1.52 The finite symmetric group S_n has order $n!$.

Proof A permutation σ of the set $\{1, \dots, n\}$ is determined completely by the way it maps the numbers amongst themselves. There are n

choices for where 1 maps to, then $n-1$ choices for where 2 goes (since we can map 2 to any of the remaining numbers except for the one we mapped 1 to), $n-2$ choices for where 3 maps to, and so on. This gives us $n(n-1)(n-2)\dots 1 = n!$ possible permutations of n objects, and so $|S_n| = n!$. $\square$

It so happens that $S_2 \cong \mathbb{Z}_2$ and $S_3 \cong D_3$. The symmetric group S_4 is isomorphic to the symmetry group of the tetrahedron. Later on we will meet Cayley's Theorem, which states that any group can be regarded as a group of permutations (although not, in general, the full symmetric group S_n for some value of n).

To investigate these permutation groups, we need a coherent and consistent notation, at least for permutations on finite sets.

One method, given that a permutation $\sigma: X \rightarrow X$ is determined completely by its action on the elements of the set X , is to represent it in the form of an array:

$$\begin{bmatrix} x_1 & x_2 & \dots & x_n \\ \sigma(x_1) & \sigma(x_2) & \dots & \sigma(x_n) \end{bmatrix}$$

The first row lists the elements of X , and the second lists their images under the action of σ . So, suppose that $\sigma \in S_5$ maps $1 \mapsto 1$, $2 \mapsto 3$, $3 \mapsto 5$, $4 \mapsto 4$ and $5 \mapsto 2$. Then we can represent σ as

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

The second row of this array shows the effect of applying σ to the first row. Suppose we have another permutation $\tau \in S_5$ such that $1 \mapsto 2$, $2 \mapsto 4$, $3 \mapsto 5$, $4 \mapsto 1$ and $5 \mapsto 3$. This can be represented as

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

Composition of permutations can be represented fairly simply using this notation:

$$\tau\sigma = \begin{array}{c} \sigma \\ \tau \end{array} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 5 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 3 & 1 & 4 \end{bmatrix}$$

Note Because we regard permutations as bijections, and the product operation as being composition, we write products from right to left, rather than left to right. So $\tau\sigma$ means σ followed by τ . Stacking the arrays vertically, we read down the page, so $\begin{smallmatrix} \sigma \\ \tau \end{smallmatrix}$ means σ followed by τ .

In general, composition of permutations isn't commutative, so we have to be careful of the order. For example,

$$\sigma\tau = \begin{array}{c} \tau \\ \sigma \end{array} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 5 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 2 & 1 & 5 \end{bmatrix} \neq \tau\sigma.$$

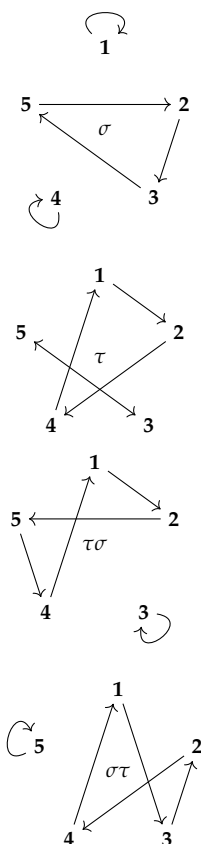


Figure 1.12: Graphical depictions of permutations σ , τ , $\tau\sigma$ and $\sigma\tau$ in S_5

This notation is quite clear, and certainly makes it easy to work out the composite of two permutations, but unfortunately it becomes somewhat unwieldy with large numbers of permuting elements. Also, it doesn't tell us very much about the actual structure of the permutation.

For example: the permutation σ maps $2 \mapsto 3$, $3 \mapsto 5$ and $5 \mapsto 2$, but leaves 1 and 4 unchanged. Repeated applications of σ cause the elements 2, 3 and 5 to cycle amongst themselves, while 1 and 4 stay where they are. Furthermore, if we apply σ three times, we get the identity permutation ι . Hence $\sigma^3 = \iota$; equivalently, σ has order 3 in S_5 .

The permutation τ , on the other hand, maps $1 \mapsto 2$, $2 \mapsto 4$ and $4 \mapsto 1$, but additionally swaps $3 \leftrightarrow 5$. So τ is doing two different, independent things at the same time: cycling $1 \mapsto 2 \mapsto 4 \mapsto 1$ and transposing $3 \leftrightarrow 5$. Neither of these two operations interfere with each other, because they are acting on disjoint subsets of the set $\{1, 2, 3, 4, 5\}$. If we apply τ three times, then the subset $\{1, 2, 4\}$ will be back where it started, but meanwhile 3 and 5 will have swapped places three times, and will hence be the other way round from where they started. But if we apply τ another three times, $\{1, 2, 4\}$ will go round another period-three cycle, and 3 and 5 will have swapped back to their original places. More concisely, $\tau^6 = \iota$.

We can depict this situation graphically, as shown in Figure 1.12, but this also isn't very compact, and tends to become more involved for larger numbers of permuting objects.

Ideally, we would like a new notation which is at the same time more compact, and also makes it easier to see at a glance the internal structure of the permutation. The key is to split the permutation into disjoint cyclic subpermutations, and write them as parenthesised lists.

Fortunately, there is a relatively straightforward procedure for decomposing a permutation as a product of disjoint cycles:

Algorithm 1.53

- Open parenthesis (
- Start with the first element 1 and write it down.
- Next write down the image $\sigma(1)$ of 1.
- Next write down the image $\sigma(\sigma(1)) = \sigma^2(1)$.
- $\vdots$
- When we get back to 1, close the parentheses).

Now repeat this process, starting with the smallest number not yet seen, and so on until you have a list of parenthesised lists of numbers. In practice, we delete any single-number lists, because they just tell us that the number in question isn't changed by the permutation.

This procedure is better illustrated with an example, so let's do it with our permutation σ above.

Starting with 1, we see that $\sigma(1) = 1$, so that gives us our first (trivial) cycle (1). Next we look at 2, which maps to 3, which maps to 5, which maps back to 2. This gives us the period-3 cycle (235). Next we look at 4, the smallest number not yet seen, which is again left unchanged by σ . This gives us another trivial cycle (4). And now we've dealt with all of the numbers in $\{1, 2, 3, 4, 5\}$, so that's the end, and our permutation can therefore be written as $(1)(235)(4)$. Except that we're really only interested in the nontrivial bits of the permutation, which in this case is the 3-cycle (235); the 1-cycles (1) and (4) don't tell us anything nontrivial about the permutation, so we discard them, leaving us with $\sigma = (235)$. This compact notation tells us that σ causes the numbers 2, 3 and 5 to cycle amongst themselves, and doesn't do anything to the remaining numbers 1 and 4.

Now let's try this with τ . This is slightly more interesting, and the procedure described above gives us the cycle notation $(124)(35)$. Just by looking at this, we can immediately see what the permutation does: it cycles $1 \mapsto 2 \mapsto 4 \mapsto 1$ and at the same time swaps $3 \leftrightarrow 5$.

The above argument tells us the following fact:

Proposition 1.54 *Any permutation $\sigma \in S_n$ can be written as a product of disjoint cyclic permutations.*

Because these cyclic permutations are disjoint, they are independent: they don't interact with each other, in the sense that since they act on disjoint subsets of $X = \{1, 2, 3, 4, 5\}$, they don't step on each others' toes. Therefore, it doesn't matter what order we apply them to the set X , and so the cycles (124) and (35) commute. (Non-disjoint cycles don't commute, however.)

A slight disadvantage with cycle notation is that it isn't so obvious how to multiply permutations together. But with a little practice it's easier than it might appear, and quite possible to do it in your head.

Using the permutations $\sigma = (235)$ and $\tau = (124)(35)$ as before, we calculate $\sigma\tau$ as follows:

$$\sigma\tau = (235)(124)(35)$$

Start with 1, and read through the list of cycles from right to left, applying each one in turn until you've done them all:

$$1 \xrightarrow{(35)} 1 \xrightarrow{(124)} 2 \xrightarrow{(235)} 3$$

Now do the same process, but starting with the number (in this case 3)

that you ended up with last time:

$$3 \xrightarrow{(35)} 5 \xrightarrow{(124)} 5 \xrightarrow{(235)} 2$$

Now do it again, and again, until you end up back at 1:

$$2 \xrightarrow{(35)} 2 \xrightarrow{(124)} 4 \xrightarrow{(235)} 4$$

$$4 \xrightarrow{(35)} 4 \xrightarrow{(124)} 1 \xrightarrow{(235)} 1$$

This gives the first disjoint cycle (1324) . Now repeat this process with the smallest number (in this case 5) not yet seen:

$$5 \xrightarrow{(35)} 3 \xrightarrow{(124)} 3 \xrightarrow{(235)} 5$$

Thus 5 is unchanged by the product $\sigma\tau$, so our next cycle is (5) , except that by convention we omit length-1 cycles. Since all of the numbers $1, \dots, 5$ are now accounted for, we are done, and $\sigma\tau = (1324)$.

Now try it with $\tau\sigma = (124)(35)(235)$.

$$\left. \begin{array}{cccc} 1 & \xrightarrow{(235)} & 1 & \xrightarrow{(35)} & 1 & \xrightarrow{(124)} & 2 \\ 2 & \xrightarrow{\quad} & 3 & \xrightarrow{\quad} & 5 & \xrightarrow{\quad} & 5 \\ 5 & \xrightarrow{\quad} & 2 & \xrightarrow{\quad} & 2 & \xrightarrow{\quad} & 4 \\ 4 & \xrightarrow{\quad} & 4 & \xrightarrow{\quad} & 4 & \xrightarrow{\quad} & 1 \end{array} \right\} (1254)$$

$$\left. \begin{array}{cccc} 3 & \xrightarrow{(235)} & 5 & \xrightarrow{(35)} & 3 & \xrightarrow{(124)} & 3 \end{array} \right\} (3)$$

So $\tau\sigma = (1254)$.

Now we introduce a couple of concepts relating to cyclic permutations which will become useful in the next bit.

Definition 1.55 Let $\sigma = (x_1 x_2 \dots x_k)$ be a finite cyclic permutation in some (possibly infinite) symmetric group $\text{Sym}(X)$. Then σ has **length** or **periodicity** k , which we denote by $l(\sigma) = k$. This is equal to the order $|\sigma|$ of σ in the group $\text{Sym}(X)$.

Cycles of length 2 will play an important rôle in the next part of the discussion, so we give them a special name:

Definition 1.56 Let $\sigma = (x_1 x_2)$ be a period-2 cyclic permutation in some (possibly infinite) symmetric group $\text{Sym}(X)$. Then we call σ a **transposition**.

We noted earlier that $\sigma^3 = \iota$ and $\tau^6 = \iota$. More formally, the order $|\sigma|$ of σ in S_5 is 3, while the order $|\tau|$ of τ in S_5 is 6. The next proposition shows that we can easily work out the order of a permutation $\pi \in S_n$ just by looking at its disjoint cycle representation.

Proposition 1.57 Let $\pi \in S_n$ be a permutation in some (possibly infinite) symmetric group $\text{Sym}(X)$. Then if π can be represented as a finite product

$$\pi = \sigma_1 \sigma_2 \dots \sigma_k$$

of disjoint cycles, each of length $l_i = l(\sigma_i)$, then

$$|\pi| = \text{lcm}(l_1, \dots, l_k).$$

That is, the order of π is the lowest common multiple of the lengths of the disjoint cycles $\sigma_1, \dots, \sigma_k$.

Proof First, observe that $\pi^n = (\sigma_1 \dots \sigma_k)^n = \sigma_1^n \dots \sigma_k^n$ since disjoint cycles commute. Then if $|\pi| = n$ this means that $\pi^n = \iota$ and hence $\sigma_1^n \dots \sigma_k^n = \iota$.

Now suppose that $\sigma_i = (x_1 \dots x_{l_i})$. Then $\pi^n = \iota$ implies that $\pi^n(x_j) = x_j$ for $1 \leq j \leq l_i$. So $(\sigma_1^n \dots \sigma_k^n)(x_j) = x_j$, and since the σ s are disjoint cycles, we know that no other cycle apart from σ_i affects x_j and therefore $\sigma_i(x_j) = x_j$ for $1 \leq j \leq l_i$. Hence $\sigma_i^n = 1$ and hence n must either equal $l_i = |\sigma_i|$ or be an integer multiple of it. That is, $l_i | n$.

Repeating this argument for all the cycles $\sigma_1, \dots, \sigma_k$ we see that if $\pi^n = \iota$ then $l_i | n$ for $1 \leq i \leq k$.

Conversely, suppose that n is such that $l_i | n$ for $1 \leq i \leq k$. Then $\pi^n = \sigma_1^n \dots \sigma_k^n = \iota \dots \iota = \iota$.

Therefore, $\pi^n = \iota$ if and only if $l_i | n$ for $1 \leq i \leq k$. So in particular, n must be the *smallest* positive integer which divides each of the l_i , which is to say that it must be the lowest common multiple $\text{lcm}(l_1, \dots, l_k)$. $\square$

So $|\sigma| = l(\sigma) = 3$, but $|\tau| = \text{lcm}(2, 3) = 6$. Now let's look again at the permutation $\sigma = (235)$. This has length 3, but is it possible to express it as a product of even shorter (but possibly not disjoint) cyclic permutations? Yes it is, and $(25)(23)$ is one possible decomposition:

$$\begin{array}{ccccc} 2 & \xrightarrow{(23)} & 3 & \xrightarrow{(25)} & 3 \\ 3 & \mapsto & 2 & \mapsto & 5 \\ 5 & \mapsto & 5 & \mapsto & 2 \end{array}$$

In general, we have the following proposition:

Proposition 1.58 Any finite permutation σ in some (possibly infinite) symmetric group $\text{Sym}(X)$ can be expressed as a product of (not necessarily disjoint) transpositions.

Proof Any finite permutation $\sigma \in \text{Sym}(X)$ can be written as a product of disjoint cyclic permutations. Furthermore, any cyclic permutation $(x_1 x_2 \dots x_k)$ can be written as a product

$$(x_1 x_2 \dots x_k) = (x_1 x_k)(x_1 x_{k-1}) \dots (x_1 x_3)(x_1 x_2)$$

of transpositions. $\square$

Corollary 1.59 The transpositions in S_n generate S_n .

In general, there may be multiple ways to represent a given permutation as a product of transpositions. The proof of Proposition 1.58

gives one valid way, but since any cyclic permutation $(x_1 x_2 \dots x_k)$ can be rewritten as, for example, $(x_k x_1 x_2 \dots x_{k-1})$ (or any other cyclic permutation of that list), the decomposition will not be unique.

Proposition 1.60 *The symmetric group S_n is generated by any of the following sets of cyclic permutations:*

- (i) $\{(12), (13), \dots, (1n)\}$
- (ii) $\{(12), (23), \dots, (n-1n)\}$
- (iii) $\{(12), (12 \dots n)\}$

Proof (i) Corollary 1.59 tells us that S_n is generated by transpositions. But any transposition (ab) can be written as a product $(1a)(1b)(1a)$.

(ii) This follows from part (i) together with the observation that $(1k) = (k-1k) \dots (34)(23)(12)(23)(34) \dots (k-1k)$.

(iii) We can write any transposition of the form $(kk+1)$ as the product $(12 \dots n)^{k-1}(12)(12 \dots n)^{1-k}$, and hence by part (ii), the transposition (12) together with the n -cycle $(12 \dots n)$ generate all of S_n .

This completes the proof. $\square$

As remarked above, the decomposition of a permutation σ into transpositions will not, in general, be unique, nor need those transpositions be disjoint. It need not even be the case that two such decompositions have the same *number* of transpositions; however we can at least say that the *parity* (odd or even) of the number of transpositions will be the same for any such decompositions of the same permutation.

Proposition 1.61 *Let σ be a finite permutation in some finite symmetric group S_n . Let $\sigma = (x_1 x_2)(x_3 x_4) \dots (x_{2k-1} x_{2k})$ be a decomposition of σ into k transpositions, and let $\sigma = (y_1 y_2)(y_3 y_4) \dots (y_{2h-1} y_{2h})$ be another decomposition of σ into h transpositions. Then $h \equiv k \pmod{2}$.*

Proof Consider the polynomial

$$\begin{aligned} P(x_1, \dots, x_n) &= \prod_{1 \leq i < j \leq n} (x_i - x_j) \\ &= (x_1 - x_2)(x_1 - x_3) \dots (x_1 - x_n)(x_2 - x_3) \dots (x_{n-1} - x_n). \end{aligned}$$

Given a permutation $\sigma \in S_n$, we define

$$\sigma(P)(x_1, \dots, x_n) = P(x_{\sigma(1)}, \dots, x_{\sigma(n)}).$$

The permuted polynomial $\sigma(P)(x_1, \dots, x_n)$ has exactly the same factors $(x_i - x_j)$ as $P(x_1, \dots, x_n)$, except that some of the variables x_i have been permuted, and some of the factors will change sign. We may therefore define the **sign** of σ to be the quotient

$$\text{sign}(\sigma) = \frac{\sigma(P)(x_1, \dots, x_n)}{P(x_1, \dots, x_n)} = \pm 1.$$

In general, for any two permutations $\sigma, \tau \in S_n$,

$$\begin{aligned}
\text{sign}(\sigma\tau) &= \frac{\sigma\tau(P)(x_1, \dots, x_n)}{P(x_1, \dots, x_n)} = \frac{P(x_{\sigma\tau(1)}, \dots, x_{\sigma\tau(n)})}{P(x_1, \dots, x_n)} \\
&= \frac{P(x_{\sigma(1)}, \dots, x_{\sigma(n)})}{P(x_1, \dots, x_n)} \frac{P(x_{\sigma\tau(1)}, \dots, x_{\sigma\tau(n)})}{P(x_{\sigma(1)}, \dots, x_{\sigma(n)})} \\
&= \frac{P(x_{\sigma(1)}, \dots, x_{\sigma(n)})}{P(x_1, \dots, x_n)} \frac{P(x_{\tau(1)}, \dots, x_{\tau(n)})}{P(x_1, \dots, x_n)} \\
&= \text{sign}(\sigma) \text{sign}(\tau).
\end{aligned}$$

The transposition $(1\ 2)$ has negative sign, since

$$\text{sign}(1\ 2) = \frac{P(x_2, x_1, \dots, x_n)}{P(x_1, x_2, \dots, x_n)} = \frac{x_2 - x_1}{x_1 - x_2} = -1.$$

In fact, all transpositions have negative sign, since part (i) of Proposition 1.60 tells us that a transposition $\tau = (a\ b) \in S_n$ can be written as a product $(1\ a)(1\ b)(1\ a)$, which also has sign -1 . Therefore, by the multiplicativity condition above, if $\sigma \in S_n$ can be decomposed as a product of an even number of transpositions, then $\text{sign}(\sigma) = +1$, and if it decomposes as an odd number of transpositions, then $\text{sign}(\sigma) = -1$.

Since the sign of a permutation is independent of any particular decomposition, a permutation with positive sign must be equal to a product of an even number of transpositions, while a permutation with negative sign must be a product of an odd number of transpositions.

So, if $\sigma = (x_1\ x_2) \dots (x_{2k-1}\ x_{2k}) = (y_1\ y_2) \dots (y_{2h-1}\ y_{2h})$, then h and k must be either both odd or both even; that is, $h \equiv k \pmod{2}$. $\square$

For example, let $\sigma = (1\ 3\ 4) = (1\ 3)(1\ 4) \in S_4$. Then

$$P(x_1, x_2, x_3, x_4) = (x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4)$$

and hence

$$\begin{aligned}
(\sigma P)(x_1, x_2, x_3, x_4) &= (x_{\sigma(1)} - x_{\sigma(2)})(x_{\sigma(1)} - x_{\sigma(3)})(x_{\sigma(1)} - x_{\sigma(4)}) \\
&\quad (x_{\sigma(2)} - x_{\sigma(3)})(x_{\sigma(2)} - x_{\sigma(4)})(x_{\sigma(3)} - x_{\sigma(4)}) \\
&= (x_3 - x_2)(x_3 - x_4)(x_3 - x_1)(x_2 - x_4)(x_2 - x_1)(x_4 - x_1) \\
&= P(x_1, x_2, x_3, x_4)
\end{aligned}$$

On the other hand, consider the permutation $\tau = (1\ 2) \in S_4$. Then

$$\begin{aligned}
(\tau P)(x_1, x_2, x_3, x_4) &= (x_{\tau(1)} - x_{\tau(2)})(x_{\tau(1)} - x_{\tau(3)})(x_{\tau(1)} - x_{\tau(4)}) \\
&\quad (x_{\tau(2)} - x_{\tau(3)})(x_{\tau(2)} - x_{\tau(4)})(x_{\tau(3)} - x_{\tau(4)}) \\
&= (x_2 - x_1)(x_2 - x_3)(x_2 - x_4)(x_1 - x_3)(x_1 - x_4)(x_3 - x_4) \\
&= -P(x_1, x_2, x_3, x_4)
\end{aligned}$$

Definition 1.62 Let $\sigma \in S_n$ be a finite permutation. If σ decomposes into an even number of transpositions (that is, $\text{sign}(\sigma) = +1$) we say it is an **even** permutation. Otherwise we call it an **odd** permutation.

The product of two odd permutations or two even permutations is clearly going to be an even permutation, while the product of an odd permutation with an even one is clearly going to be odd. Furthermore, the parity (odd or even) of an inverse permutation σ^{-1} will be equal to the parity of the original permutation σ . Finally, the identity permutation ι is even (because it decomposes to a product of zero permutations, and zero is even). This gives another family of groups related to the symmetric groups S_n :

Example 1.63 The **alternating group** A_n is the permutation group that consists of all even permutations of the set $\{1, \dots, n\}$. This set is closed under composition of permutations (and hence composition is a well-defined binary operation on A_n), it contains the identity permutation ι , and every even permutation σ has a unique inverse σ^{-1} which is also even, by the remarks above.

The order of A_n is $n!/2$, half the order of the symmetric group S_n , since exactly half of the permutations on n elements are even.

By analogy with $\text{Sym}(X)$, we may sometimes use $\text{Alt}(X)$ to denote the alternating group on a (possibly infinite) set X .

The groups A_1 and A_2 are trivial, $A_3 \cong \mathbb{Z}_3$ and A_4 is isomorphic to the rotation group of the tetrahedron. The alternating groups A_n for $n \geq 5$ are examples of what are known as **simple** groups, which we will discuss further in Chapter 3.

Just as the full symmetric group S_n is generated by transpositions, it turns out that (apart from the two trivial cases A_1 and A_2) the alternating groups are generated by the period-3 cyclic permutations:

Proposition 1.64 For $n \geq 3$, any even permutation on the set $\{1, \dots, n\}$ may be written as a product of 3-cycles.

Proof Given an element $\sigma \in A_n$, by Proposition 1.60 we may write it as the product of an even number of transpositions of the form

$$\sigma = (1\ x_1)(1\ x_2) \dots (1\ x_{2k}).$$

Each adjacent pair of transpositions $(1\ x_{2i-1})(1\ x_{2i})$ may now be rewritten as a 3-cycle $(1\ x_{2i-1}\ x_{2i})$. □

This completes our first venture into group theory, and also our first look at symmetric groups. In the next chapter we will delve further into the subject, studying **subgroups**: groups within groups. The importance of permutation groups will become clear: by Cayley's Theorem,¹⁰ every group is isomorphic to a permutation group.

¹⁰ Theorem 2.13, page 47.

Summary

In this chapter we began our foray into abstract algebra by distilling the basic properties of integer addition, probably the first number system any of us meet. We found that this number system $\mathbb{Z}$ consisted of a set $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ equipped with a **binary operation**¹¹ satisfying certain properties: **commutativity**¹², **associativity**¹³, existence of an **identity element**¹⁴, and existence of **inverse elements**.¹⁵

¹¹ Definition 1.1, page 2.

¹² Definition 1.3, page 3.

¹³ Definition 1.4, page 3.

¹⁴ Definition 1.6, page 4.

¹⁵ Definition 1.8, page 6.

Restating this in more general terms we obtain the concept of a **group**:¹⁶ a set G equipped with an associative binary operation $*$ such that G contains a special identity element e , and each element g of G has an inverse g^{-1} . A group whose operation $*$ is commutative as well is said to be an **abelian group**.¹⁷ The **order**¹⁸ $|G|$ of a group G is the number of elements it has.

¹⁶ Definition 1.9, page 6.

¹⁷ Definition 1.11, page 7.

¹⁸ Definition 1.10, page 6.

Certain consequences follow more or less immediately from this definition: the identity element e is unique in a given group G (that is, there is no other element f satisfying the required properties), as is the inverse g^{-1} of any given element g .¹⁹ The left and right cancellation laws also follow from the definition.²⁰ The inverse of a product of two elements is equal to the product of the corresponding inverses, but written in the opposite order: $(g*h)^{-1} = h^{-1}*g^{-1}$.

¹⁹ Proposition 1.16, page 8.

²⁰ Proposition 1.15, page 8.

Probably the simplest examples of groups apart from $\mathbb{Z}$ itself and the **trivial group**²¹ $\{e\}$ are the finite **cyclic groups**²² $\mathbb{Z}_n = \{0, 1, \dots, n-1\}$ with modulo- n addition. A simple way of studying the internal structure of a group is by examining its **multiplication table**. The table for $\mathbb{Z}_3$, suitably relabelled, is structurally identical to that for the set $\{1, \omega, \omega^2\}$ of complex cube roots of unity. This leads us to the notion of an **isomorphism**:²³ a bijection from one group to another which respects the group structures. More precisely, a bijection $f: G \rightarrow H$ is an isomorphism if $f(g_1*g_2) = f(g_1)\circ f(g_2)$ for all $g_1, g_2 \in G$.

²¹ Example 1.13, page 7.

²² Example 1.12, page 7.

²³ Definition 1.17, page 11.

Another tool for understanding the internal structure of a group is the **order**²⁴ $|g|$ of an element g : the smallest natural number n such that $g^n = e$ (or $ng = e$ if we're using additive notation). If no such n exists, we say the element is of infinite order. The order $|e|$ of the identity element is 1, and indeed e is the only element with order 1.²⁵ If every element of G apart from e has order 2, then G must be abelian.²⁶ The order of an element is invariant under isomorphisms: the image of an element has the same order in its group as the element itself did in the original group.²⁷

²⁴ Definition 1.18, page 11.

²⁵ Proposition 1.19, page 11.

²⁶ Proposition 1.20, page 11.

²⁷ Proposition 1.21, page 12.

Some groups can be reconstructed from a proper subset of elements of the group. We call those elements **generators** and, if there are finitely

²⁸ Definition 1.22, page 12.

²⁹ Definition 1.23, page 12.

³⁰ Proposition 1.24, page 13.

³¹ Proposition 1.25, page 13.

³² Proposition 1.26, page 13.

³³ Definition 1.27, page 14.

³⁴ Proposition 1.28, page 14.

³⁵ Example 1.30, page 15.

³⁶ Proposition 1.32, page 16.

³⁷ Example 1.34, page 17.

³⁸ Example 1.36, page 18.

³⁹ Example 1.38, page 18.

⁴⁰ Example 1.35, page 17.

⁴¹ Example 1.37, page 18.

⁴² Example 1.39, page 18.

⁴³ Example 1.40, page 18.

⁴⁴ Example 1.42, page 19.

⁴⁵ Definition 1.46, page 22.

⁴⁶ Definition 1.43, page 21.

⁴⁷ Example 1.45, page 21.

⁴⁸ Definition 1.47, page 22.

⁴⁹ Definition 1.48, page 22.

⁵⁰ Definition 1.50, page 24.

many of them, say that the group is **finitely generated**.²⁸ In the case of the integers $\mathbb{Z}$ or the finite cyclic groups $\mathbb{Z}_n$, we can recover the whole of the group by taking finite products (or sums) of a single nontrivial generator; groups which are generated by a single element are said to be **cyclic**.²⁹ There is only one cyclic group (up to isomorphism) of a given (**finite** or **infinite**) order.³⁰ The order of a (finite or infinite) cyclic group is equal to the order of any of its generators.³¹ For the finite cyclic groups $\mathbb{Z}_n$, a nonzero integer $k \in \mathbb{Z}_n$ generates the whole group if and only if k and n are coprime.³²

Given two groups G and H we can form a new group by taking the cartesian product $G \times H$ and defining the obvious combined group structure on it; we call this the **direct product** $G \times H$ (or, if we're using additive notation, the **direct sum** $G \oplus H$).^{33,34} In particular, the Klein 4-group V_4 is isomorphic to the direct sum $\mathbb{Z}_2 \oplus \mathbb{Z}_2$,³⁵ and in general $\mathbb{Z}_m \oplus \mathbb{Z}_n \cong \mathbb{Z}_{mn}$ if and only if m and n are coprime.³⁶

Matrices yield an important and rich class of groups, many of which are nonabelian. Among these are the **general linear groups** $GL_n(\mathbb{K})$ of invertible matrices,³⁷ the **orthogonal groups** $O_n(\mathbb{K})$ of matrices whose inverse is equal to their transpose,³⁸ and the **unitary groups** of complex matrices whose inverse is equal to their conjugate transpose.³⁹ (Here $\mathbb{K}$ can be either $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$, and later we will consider the case of a **finite field** $\mathbb{F}_p$.) Each of these groups has a **special** version consisting of just the matrices of the given type with determinant 1,^{40,41,42} and in addition we have the **symplectic groups** $Sp_{2n}(\mathbb{K})$ of **symplectic matrices**.⁴³ An interesting and important finite example is given by Pauli's **spin matrices**; this is closely related to the eight-element **quaternion group** Q_8 .⁴⁴

Matrix groups led naturally to a discussion of geometric operations on vector spaces, and in particular we considered **isometries**: operations which preserve distance. Isometries which preserve orientation are said to be **direct**, while those which reverse orientation are said to be **opposite**.⁴⁵ Considering those isometries which map a particular geometric object to itself leads to the notion of symmetry groups such as the **dihedral groups** D_n : these comprise the $2n$ symmetry operations (reflections and rotations) we can perform on a regular n -gon.⁴⁶ The Klein 4-group V_4 can similarly be viewed as the symmetry group of a rectangle.⁴⁷ More generally, the **Euclidean groups** E_n^+ and E_n are defined as, respectively, the direct and full isometry groups $\text{Isom}^+(\mathbb{R}^n)$ and $\text{Isom}(\mathbb{R}^n)$ of $\mathbb{R}^n$, and in particular we characterised the various types of elements in the two- and three-dimensional case.^{48,49}

In turn, symmetry groups led to a discussion of **permutations**: bijections $\sigma: X \rightarrow X$ from a set to itself.⁵⁰ Composition of permutations is

associative, there is an obvious identity permutation ι , and bijections are necessarily invertible. This leads to the notion of the **symmetric group** $\text{Sym}(X)$ on a set X ;⁵¹ in the case where X is a finite set with n elements, we usually denote this S_n (a simple argument⁵² shows that $|S_n| = n!$). We examined two different notations for finite permutations: array notation and cycle notation.⁵³ The first of these is perhaps more intuitive in some respects, and it's a little easier to compose permutations written in this way, but it doesn't really give a clear insight into the internal structure of the permutation. The latter is more compact and does allow us to see how individual objects are shuffled around by the permutation, and with a little bit of practice it becomes almost as easy to write down the composition of two permutations.

A simple argument demonstrates that any finite permutation can be written as a product of disjoint cycles; that is, cycles which don't affect the same objects, and hence don't interact with each other.⁵⁴ The order of a cycle is (reasonably obviously) equal to its length,⁵⁵ and (less obviously) the order of a product of disjoint cycles is equal to the lowest common multiple of the lengths of the individual cycles.⁵⁶ We give a cycle of length 2 (that is, a permutation which just interchanges two distinct objects and leaves everything else where it is) a special name: we call it a **transposition**.⁵⁷ It transpires that any finite permutation can be written as a product of (not necessarily disjoint) transpositions; that is, any finite permutation can be performed by carefully swapping pairs of elements in some sequence.⁵⁸ This sequence of transpositions will not, in general, be unique for a given permutation; even worse, two different such decompositions of a given permutation won't necessarily even have the same number of transpositions. However, it is at least the case that all such sequences for a given permutation will have the same **parity**: a permutation σ will always be decomposable into either an odd or an even number of transpositions.⁵⁹ This gives us a neat classification of finite permutations into two types: **odd** or **even**.⁶⁰ The even permutations of n objects form a smaller group called A_n , the **alternating group**, which is exactly half the size of the full symmetric group S_n (which contains both the odd and the even permutations).⁶¹ It is straightforward to show that any even finite permutation on three or more objects can be written as a product of length-3 cycles.⁶²

⁵¹ Definition 1.51, page 24.

⁵² Proposition 1.52, page 24.

⁵³ Algorithm 1.53, page 26.

⁵⁴ Proposition 1.54, page 27.

⁵⁵ Definition 1.55, page 28.

⁵⁶ Proposition 1.57, page 28.

⁵⁷ Definition 1.56, page 28.

⁵⁸ Proposition 1.58, page 29.

⁵⁹ Proposition 1.61, page 30.

⁶⁰ Definition 1.62, page 31.

⁶¹ Example 1.63, page 32.

⁶² Proposition 1.64, page 32.

References and further reading

- I Kleiner, *The evolution of group theory: a brief survey*, Mathematics Magazine 59.4 (1986) 195–215
- H Wussing, *The Genesis of the Abstract Group Concept*, Dover (2007)

Two overviews of the history of group theory from the late 18th century to the early 20th century.

Exercises

- 1.1 Let $*$ be a binary operation defined on $\mathbb{R}$ such that $x * y = x + y + xy$.
 - (a) Is $*$ associative?
 - (b) Is $*$ commutative?
 - (c) Is there an element $e \in \mathbb{R}$ which serves as an identity with respect to $*$?
 - (d) Does every element $x \in \mathbb{R}$ have an inverse x^{-1} with respect to $*$?
- 1.2 Let $S = \{a, b\}$ be a set with two elements. How many possible binary operations can be defined on this set? How many of these determine semigroups, that is, how many are associative? How many of those determine monoids? How many determine a group structure?
- 1.3 Which of the following binary operations (defined on the set $\mathbb{R}$ of real numbers) are associative, and which are commutative?
 - (a) $x * y = x + 2y$
 - (b) $x * y = x + y + xy + 1$
 - (c) $x * y = x$
- 1.4 Show that the multiplication table for a finite group G satisfies the **Latin square** property. That is, show that each element of the group occurs exactly once in each row and column of the table.
- 1.5 Let $*$: $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ by $x * y = xy + 1$. Is this operation commutative? Does it determine a group structure on $\mathbb{R}$? If so, prove it by verifying the necessary criteria; if not, which axioms fail?
- 1.6 Which of the following are groups? For any that are not, say which of the axioms Go–G3 fail.

(a) $(\mathbb{Z}, \times)$	(e) $(\mathbb{Q}, *)$ where $\frac{a}{b} * \frac{c}{d} = \frac{a+c}{b+d}$, for $a, b, c, d \in \mathbb{Z}$ and $b, d > 0$
(b) $(\mathbb{Z}_3, \times_3)$	(f) $(\mathbb{Z}_{\text{even}}, +)$
(c) $(\mathbb{Z}_4, \times_4)$	(g) $(\mathbb{Z}_{\text{odd}}, +)$
(d) $(\mathbb{Z}_n, \times_n)$ if n isn't prime	(h) $(\mathbb{Z}, -)$

Here, $\times$ represents the usual multiplication operation defined on $\mathbb{Z}$, and $\times_n$ represents multiplication modulo n .

- 1.7 Given a group G , and some fixed element $g \in G$, show that the function $\phi_g: x \mapsto g^{-1}xg$ is an isomorphism $G \rightarrow G$.
- 1.8 Write out the group multiplication table for $\mathbb{Z}_2 \oplus \mathbb{Z}_5$.
- 1.9 Let G be a group. Show that if $(ab)^2 = a^2b^2$ for all $a, b \in G$, then G must be abelian.
- 1.10 Let G be a nonempty set and let $*$: $G \times G \rightarrow G$ be an associative binary operation. Suppose that there is a distinguished element $e \in G$ such that:
 - (a) $e * x = x$ for all $x \in G$. (That is, e is a **left identity**.)
 - (b) For any $x \in G$, there exists $y \in G$ such that $y * x = e$. (That is, G has **left inverses**.)
 Then show that $G = (G, *)$ is a group. (This demonstrates that the conditions in Definition 1.9 are slightly stronger than necessary.)

- 1.11 Let G be a group, and suppose that $g \in G$ and $m, n \in \mathbb{Z}$. Prove that $(g^m)^n = g^{mn}$ and $g^m g^n = g^{m+n}$.
- 1.12 Show that $|a * b| = |b * a|$ for any elements a, b of some group $G = (G, *)$.
- 1.13 Show that any finite group of even order has an odd number of elements of order 2.
- 1.14 Show that $\text{Isom}(\Delta^3)$, the full isometry group of the tetrahedron, is isomorphic to the symmetric group S_4 , and that the direct isometry group $\text{Isom}^+(\Delta^3)$ is isomorphic to the alternating group A_4 .
- 1.15 Let $\sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 5 & 1 & 4 & 6 \end{bmatrix}$ be an element of S_6 . Write this permutation as a product of disjoint cycles, and then as a product of transpositions.
- 1.16 Write the following elements of S_7 as (i) a product of disjoint cycles and (ii) as a product of transpositions, and (iii) calculate their orders.
- (a) $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 5 & 4 & 1 & 7 & 2 & 6 \end{bmatrix}$ (e) $(3\ 1\ 2\ 4\ 5)(4\ 2\ 1\ 3)(2\ 1)(4\ 1\ 2)$
- (b) $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 4 & 7 & 1 & 6 & 5 & 3 \end{bmatrix}$ (f) $(2\ 7)(1\ 3\ 4)(7\ 2)(6\ 5\ 3)$
- (c) $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 1 & 4 & 3 & 5 & 7 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 6 & 2 & 1 & 7 & 3 & 5 \end{bmatrix}$
- (d) $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 7 & 6 & 3 & 5 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 2 & 6 & 3 & 5 & 1 & 7 \end{bmatrix}$
- 1.17 The 3-cycle $(1\ 2\ 3)$ has order 3, the highest possible for an element of the symmetric group S_3 ; similarly $(1\ 2\ 3\ 4)$ has maximal order 4 in S_4 , while $(1\ 2\ 3)(4\ 5)$ has order 6, which is the largest possible in S_5 . Find examples of elements of maximal order in the symmetric groups S_6 to S_{12} .
- 1.18 Show that $\text{Sym}(A)$ is nonabelian if $|A| > 2$.
- 1.19 What is the highest order of any element of S_7 ? How many permutations in S_7 have this order?
- 1.20 For $n \in \mathbb{Z}$ define a function $f_n: \mathbb{Z} \rightarrow \mathbb{Z}$ by $f_n(m) = m + n$ for any $m \in \mathbb{Z}$. Show that $f_n \in \text{Sym}(\mathbb{Z})$, and also that $f_n \circ f_m = f_{m+n} = f_m \circ f_n$ and $f_n^{-1} = f_{-n}$ for any $m, n \in \mathbb{Z}$.
- 1.21 (a) Show that the function $f: \mathbb{Z}_{13} \rightarrow \mathbb{Z}_{13}$ given by $f(x) = x^5$ is a permutation. Write this permutation in disjoint cycle notation.
- (b) More generally, choose some integer $0 < \alpha < 13$, show that the **Dickson polynomial** $D_5(x, \alpha) = x^5 - 5\alpha x^3 + 5\alpha^2 x$ defines a permutation of $\mathbb{Z}_{13}$, and write this permutation in disjoint cycle notation.
- (These are examples of **permutation polynomials**, which have applications in the design of error-correcting codes for data transmission and storage.)
- 1.22 Let $A \in GL_n(\mathbb{R})$ be some invertible $n \times n$ matrix over $\mathbb{R}$. Define a map $f_A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ by $f_A(\mathbf{v}) = A\mathbf{v}$ and show that $f_A \in \text{Sym}(\mathbb{R})$.
- 1.23 Let $S = \mathbb{R} \setminus \{-1, 0, 1\}$ and let $f: S \rightarrow S$ map $x \mapsto \frac{1+x}{1-x}$. Denote by f^n the n -fold composite $f \circ f \circ \cdots \circ f$. Find the order of f ; that is, the smallest positive integer n such that $f^n = \text{id}$, and thereby show that $\langle f \rangle \cong \mathbb{Z}_n$.
- 1.24 Show that $Sp_{2n}(\mathbb{R})$ is indeed a group.
- 1.25 Show that $GL_n(\mathbb{Z})$ is abelian if and only if $n = 1$.
- 1.26 Show that the rotation group of the cube is isomorphic to S_4 .

2 Subgroups

A CAREFUL EXAMINATION of the multiplication tables of many of the groups we met in the previous chapter reveals a wealth of internal structure. The dihedral group D_3 , for example, has an obvious block consisting of the identity and the clockwise and anticlockwise rotations. We could throw away the three reflection operations and still be left with a perfectly respectable group, which in this case happens to be the rotation group $R_3 = \text{Isom}^+(\Delta^2) \cong \mathbb{Z}_3$. So in some sense, D_3 has a copy of R_3 (or, for that matter, $\mathbb{Z}_3$) inside it.

Similarly, we could throw away everything except the identity and a single reflection (m_1 for example) and still be left with an order-2 group isomorphic to $\mathbb{Z}_2$. So, in the same sense, D_3 also has a copy (more accurately, three copies) of $\mathbb{Z}_2$ embedded in it.

Similarly, looking at the multiplication table for $\mathbb{Z}_4$ reveals an order-2 group $\{0, 2\} \cong \mathbb{Z}_2$ inside it.

In this chapter, we will make this concept more precise, and develop these ideas to learn more about the internal structure of groups. Along the way, we will prove Lagrange's Theorem, which imposes an important necessary (but not sufficient) condition on the order of these groups-within-groups, and use it to prove Euler's Theorem and Fermat's Little Theorem, two important and useful number-theoretic results about the factorisation of integers.

2.1 Groups within groups

THE MAIN CONCEPT WE WILL EXPLORE IN THIS CHAPTER is that of a **subgroup**: a smaller group neatly embedded inside a larger one. More precisely, given some group $G = (G, *)$, we are interested in finding a subset $H \subseteq G$ which is a group in its own right under the "same" binary operation $*$ as G . Not just any subset will do, however, so we need to decide what criteria we want a given subset to satisfy.

These purely mathematical sciences of algebra and geometry are sciences of the pure reason deriving no weight and no assistance from experiment, and isolated, or at least isolable, from all outward and accidental phenomena.

— William Rowan Hamilton
(1805–1865),

Introductory Lecture on Astronomy,
Dublin University Review 1 (1833)

$\cdot$	e	r	r^2	m_1	m_2	m_3
e	e	r	r^2	m_1	m_2	m_3
r	r	r^2	e	m_3	m_1	m_2
r^2	r^2	e	r	m_2	m_3	m_1
m_1	m_1	m_2	m_3	e	r	r^2
m_2	m_2	m_3	m_1	r^2	e	r
m_3	m_3	m_1	m_2	r	r^2	e

Table 2.1: The multiplication table for the dihedral group D_3 with the subgroup $\{e, r, r^2\}$ highlighted

$\cdot$	e	r	r^2	m_1	m_2	m_3
e	e	r	r^2	m_1	m_2	m_3
r	r	r^2	e	m_3	m_1	m_2
r^2	r^2	e	r	m_2	m_3	m_1
m_1	m_1	m_2	m_3	e	r	r^2
m_2	m_2	m_3	m_1	r^2	e	r
m_3	m_3	m_1	m_2	r	r^2	e

Table 2.2: The multiplication table for the dihedral group D_3 with the subgroup $\{e, m_1\}$ highlighted

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

Table 2.3: The multiplication table for the cyclic group $\mathbb{Z}_4$ with the subgroup $\{0, 2\}$ highlighted

On the other side, the general opinion has been and is that it is indeed by experience that we arrive at the mathematics, but that experience is not their proper foundation: the mind itself contributes something.

— Arthur Cayley (1821–1895),
Presidential Address to the British
Association for the Advancement of
Science (1883)

Given a group $G = (G, *)$ and a subset $H \subseteq G$, we impose on H the same binary operation $*$ as we use in G . We can do this, because every element of H is also contained in G , so for any $h_1, h_2 \in H \subseteq G$, the product $h_1 * h_2$ is well-defined, because h_1 and h_2 are also in G .

But we have to be careful not to gloss over a subtly important, if slightly pedantic, point here. The binary operation $*$ is a function $*$: $G \times G \rightarrow G$, but to define a group structure on H we actually need a function $*_H: H \times H \rightarrow H$. The obvious way of dealing with this is to define a new function by restricting the domain of $*$ to the subset $H \times H \subseteq G \times G$. This gives us a function $*|_{H \times H}: H \times H \rightarrow G$, which is nearly what we want, but not quite. It will work fine, though, if the image of $*|_{H \times H}$ is contained in H rather than just G . This requirement is exactly the same as requiring that H is closed under the action of $*|_{H \times H}$. If it is, then we can just define our binary operation $*_H$ to be the restriction $*|_{H \times H}$ with the codomain changed to be H instead of G . We call this new operation $*_H$ the **induced operation**, and from now on we'll drop the subscript H unless to do so would be ambiguous.

We now have a criterion that a subset $H \subseteq G$ must satisfy in order for H to be a subgroup of G : it must be closed under the action of the induced operation. In addition, $H = (H, *_H)$ must also satisfy the usual axioms for a group (see Definition 1.9). We get associativity (G1) for free: $*_H$ is associative since $*$ is, and merely restricting it to some closed subset of G isn't going to cause any problems in that regard.

It's also worth noting at this point that if G is abelian, then any subgroup will be too, for a similar reason: $*_H$ inherits commutativity from $*$ in the same way that it inherits associativity.

That leaves us with two remaining axioms to check: existence of an identity (G2) and existence of inverses (G3): we require $e \in H$ and also that $h^{-1} \in H$ for all $h \in H$. Actually, the presence of the identity element follows from the existence of inverses and the closure requirement, since if $h \in H$ and $h^{-1} \in H$, closure implies that $h *_H h^{-1} = e = h^{-1} *_H h$ must also be in H . One last thing we need to remember is that a group must have at least one element, so H can't be the empty set $\emptyset$. This discussion gives us the following definition.

Definition 2.1 Let $G = (G, *)$ be a group, let $H \subseteq G$ be a nonempty subset of G , and let $*_H$ be the induced operation obtained by restricting $*$ to $H \times H$. Then $H = (H, *_H)$ is a **subgroup** of G (written $H < G$) if and only if:

SG1 H is closed under the action of $*_H$. That is, for all $h_1, h_2 \in H$, the product $h_1 *_H h_2 \in H$ too.

SG2 For all $h \in H$, the inverse $h^{-1} \in H$ as well.

The remark about inheritance of commutativity also yields the following simple but relevant fact:

Proposition 2.2 *Let $G = (G, *)$ be abelian, and let $H < G$ be a subgroup of G . Then H is also abelian.*

The converse doesn't hold, however, since nonabelian groups can have abelian subgroups. For example, we noted earlier that D_3 has a subgroup isomorphic to $\mathbb{Z}_3$ which is therefore abelian, but D_3 itself isn't abelian.

This rotation subgroup $\{e, r, r^2\} < D_3$ is, as we noted earlier, exactly the group $\text{Isom}^+(\Delta^2)$ of direct isometries of an equilateral triangle Δ^2 , embedded neatly inside the full isometry group $D_3 = \text{Isom}(\Delta^2)$. Generalising this observation gives the following example:

Example 2.3 Let X be some geometric object (such as a subset of $\mathbb{R}^n$ for some n). The group $\text{Isom}^+(X)$ of direct isometries of X is a subgroup of the full isometry group $\text{Isom}(X)$.

The rotations of a regular n -gon form an order- n subgroup of the dihedral group D_n which is isomorphic to $\mathbb{Z}_n$. Similarly, the group $\text{Isom}^+(\Delta^3)$ from Example 1.49 is a subgroup of the full isometry group $\text{Isom}(\Delta^3)$, and the groups $E_2^+ = \text{Isom}^+(\mathbb{R}^2)$ and $E_3^+ = \text{Isom}^+(\mathbb{R}^3)$ (see Definitions 1.47 and 1.48) of, respectively, the orientation-preserving isometries of $\mathbb{R}^2$ and $\mathbb{R}^3$ are subgroups of, respectively, the full isometry groups $E_2 = \text{Isom}(\mathbb{R}^2)$ and $E_3 = \text{Isom}(\mathbb{R}^3)$.

Any nontrivial group G has at least two subgroups: the **trivial** subgroup $\{e\} < G$, and G itself. In the next section, we will meet Lagrange's Theorem, which places an elegant necessary (but not sufficient) condition on what other subgroups a given group might have.

Example 2.4 The multiplicative group $\mathbb{R}_*^\times$ of nonzero real numbers can be regarded as a subgroup of $\mathbb{C}_*^\times$, the multiplicative group of nonzero complex numbers: $\mathbb{R}_*^\times$ is clearly a subset of $\mathbb{C}_*^\times$, it's closed under multiplication (since the product of any two real numbers is also a real number) and every nonzero real number has a unique multiplicative inverse.

This last example suggests an interesting alternative viewpoint. In this chapter so far, we've chosen a group G and then looked at what smaller groups H might be contained within it. But, historically at least, the invention (or discovery, depending on one's philosophical perspective) of the various number systems $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ happened in the reverse direction: we started with one number system and then extended it in a consistent way to obtain a new number system that satisfied certain desirable properties that were lacking in

the previous one. What happens if we take a group G and, instead of examining its subgroups, we try to extend it in some way to obtain a larger group containing G as a subgroup? This turns out to be an interesting and rich line of inquiry, and we will return to it later when we look at group extensions, semidirect products and extension fields.

Example 2.5 The subset $C = \{z \in \mathbb{C} : |z| = 1\} = \{e^{i\theta} : 0 \leq \theta < 2\pi\}$, which we can regard geometrically as the unit circle in the complex plane, is a subgroup of $\mathbb{C}^\times$. It's closed under multiplication, since if $0 \leq \theta, \phi < 2\pi$, then $e^{i\theta}e^{i\phi} = e^{i(\theta+\phi)} \in C$, and it contains inverses, since if $0 \leq \theta < 2\pi$, then $e^{i\theta}e^{i(2\pi-\theta)} = e^{2\pi i} = 1$.

In fact, this subgroup is isomorphic to the unitary group U_1 . Recall that U_1 is the group of 1×1 invertible matrices A such that $A^{-1} = A^\dagger$. Unpacking this definition further, for a 1×1 complex, nonzero matrix $A = [z]$ to satisfy this requirement is equivalent to the condition $\frac{1}{z} = \bar{z}$, or $z\bar{z} = 1$. But $z\bar{z} = |z|^2$, hence $U_1 = \{[z] \in GL_1(\mathbb{C}) : |z| = 1\}$. The map $f: U_1 \rightarrow C$ given by $[z] \mapsto z$ is clearly an isomorphism.

The mention of U_1 in this example suggests that some of the other matrix groups we've encountered might be subgroups of each other.

Example 2.6 Let $\mathbb{K}$ denote one of the number systems $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$. Then the special linear group $SL_n(\mathbb{K})$, the orthogonal group $O_n(\mathbb{K})$ and the special orthogonal group $SO_n(\mathbb{K})$ are all subgroups of the general linear group $GL_n(\mathbb{K})$. Also, $SO_n(\mathbb{K}) < O_n(\mathbb{K})$.

Furthermore, the unitary group U_n and the special unitary group SU_n are both subgroups of $GL_n(\mathbb{C})$, and $SU_n < U_n$.

This example provides us with an interesting collection of groups with complicated interrelationships, which we can depict graphically by means of the **lattice diagram** or **Hasse diagram** shown in Figure 2.1.

In such a diagram, following a line up the page represents the relationship "is a subgroup of". We can do this for some of the other examples we've encountered so far, with variously illuminating results. Figure 2.2, for example, highlights the differences in internal structure between the cyclic group $\mathbb{Z}_4$ and the Klein group V_4 (see Example 1.45), while Figure 2.3 depicts the lattice diagram for D_3 .

Looking at the lattice diagram for $\mathbb{Z}_4$, we see that the subgroup $\{0, 2\}$ can actually be regarded as the cyclic group $\langle 2 \rangle$. More interestingly, the cyclic group $\mathbb{Z}_{12}$ contains four proper, nontrivial subgroups which can be generated by a single element, namely

$$\begin{aligned} \langle 2 \rangle &= \{0, 2, 4, 6, 8, 10\} & \langle 3 \rangle &= \{0, 3, 6, 9\} \\ \langle 4 \rangle &= \{0, 4, 8\} & \langle 6 \rangle &= \{0, 6\} \end{aligned}$$

which fit together in the lattice diagram shown in Figure 2.4

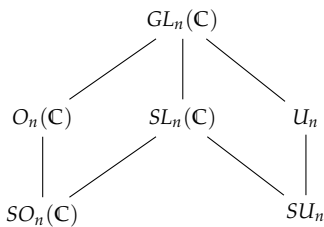


Figure 2.1: Lattice diagram depicting some subgroups of $GL_n(\mathbb{C})$

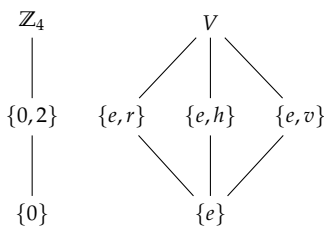


Figure 2.2: Lattice diagrams for $\mathbb{Z}_4$ and the Klein group V

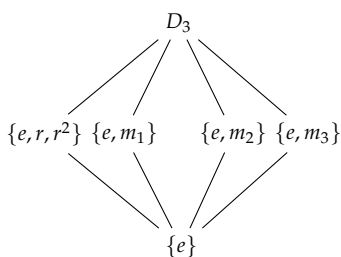


Figure 2.3: Lattice diagram for the dihedral group D_3

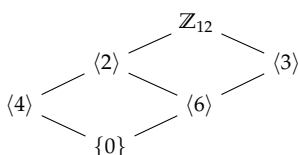


Figure 2.4: Lattice diagram for the cyclic group $\mathbb{Z}_{12}$

More generally, we have the following fact:

Proposition 2.7 *Let G be a group. Then for any $g \in G$, the set*

$$\langle g \rangle = \{g^n : n \in \mathbb{Z}\} \subseteq G$$

*forms a subgroup of G , the **cyclic subgroup** generated by the element g .*

Proof The subset $\langle g \rangle$ is clearly a subset of G . It is closed under the induced operation, since $g^m * g^n = g^{m+n} \in \langle g \rangle$ for any $m, n \in \mathbb{Z}$. And for any $g^n \in \langle g \rangle$, the inverse $(g^n)^{-1} = g^{-n}$ is also contained in $\langle g \rangle$. Hence $\langle g \rangle < G$. $\square$

Looking more carefully at the lattice diagrams for $\mathbb{Z}_4$ and $\mathbb{Z}_{12}$, we see that all of their subgroups are cyclic; we can even write $\mathbb{Z}_4 = \langle 1 \rangle$, $\mathbb{Z}_{12} = \langle 1 \rangle$ and $\{0\} = \langle 0 \rangle$ to complete the picture.

The lattice diagram for $\mathbb{Z}$ is similarly illuminating, although infinite in size. An illustrative fragment is shown in Figure 2.5.

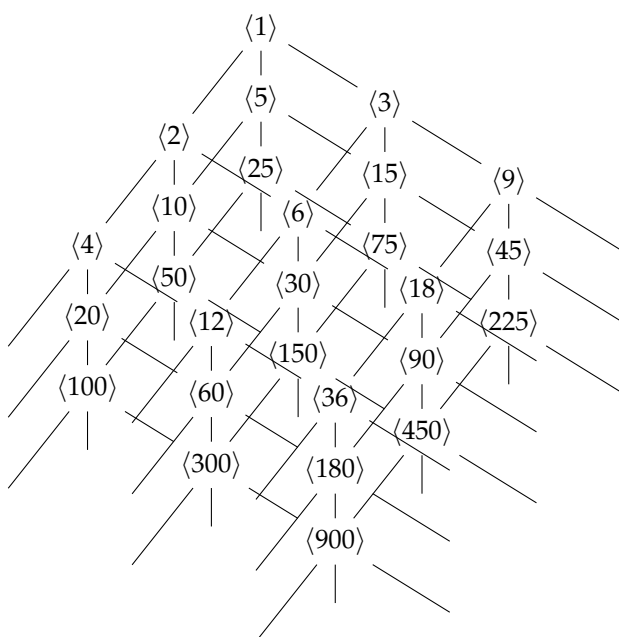


Figure 2.5: Part of the lattice diagram for subgroups of $\mathbb{Z}$

Again, all of the subgroups are cyclic, generated by some single element of $\mathbb{Z}$. (The trivial subgroup isn't, and can't be, shown: it appears as the infinitely distant bottom vertex of the lattice, contained in all of the other subgroups.) We will often write the set

$$\langle n \rangle = \{\dots, -3n, -2n, -n, 0, n, 2n, 3n, \dots\}$$

of integer multiples of n as $n\mathbb{Z}$.

These observations lead us to ask whether all subgroups of (finite or

infinite) cyclic groups are themselves cyclic, and it turns out that the answer is yes:

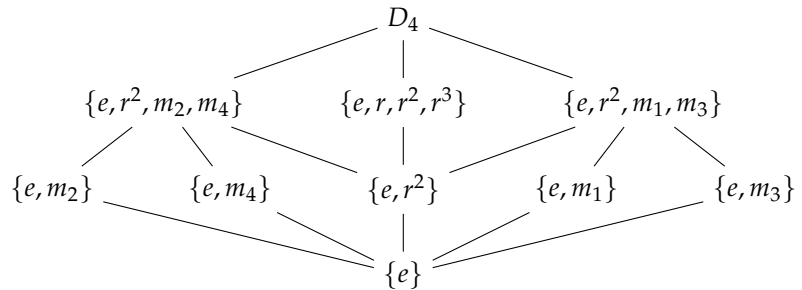
Proposition 2.8 *Let $G = \langle g \rangle$ be a (finite or infinite) cyclic group. If $H < G$ then H is also cyclic.*

Proof If H is trivial then it is cyclic, generated trivially by the identity $e \in G$. If $H = G$ then it is obviously cyclic, since G is.

If H is a proper and nontrivial subgroup of G , then choose k to be the smallest positive integer such that $g^k \in H$, where g is the generator of G . Now choose some other element g^n which also lies in H . Dividing n by k we get $n = kq + r$ for some unique integers $q, r \in \mathbb{Z}$, with $0 \leq r < k$.¹ So $g^n = g^{kq+r} = g^{kq} * g^r$, which implies that $g^r = g^{-kq} * g^n$.

Since H is a subgroup of G , it must be closed under the induced operation, and so g^r must also lie in H . So $r = 0$, otherwise k wasn't the smallest positive integer for which $g^k \in H$. Hence $g^n = g^{kq} = (g^k)^q$ and so any element of H can be written as a power of g^k . Therefore g^k generates all of H and so $H = \langle g^k \rangle$ is cyclic. $\square$

A greater wealth of structure can be seen in the lattice diagram for the dihedral group D_4 :



The subgroups on the second level are all of order 4; two of them ($\{e, r^2, m_2, m_4\}$ and $\{e, r^2, m_1, m_3\}$) are isomorphic to the Klein group V_4 , while the other (the rotation subgroup $\{e, r, r^2, r^3\}$) is isomorphic to $\mathbb{Z}_4$. The subgroups on the third level are all isomorphic to $\mathbb{Z}_2$.

For any dihedral group D_n , the rotation subgroup $\{e, r, r^2, \dots, r^{n-1}\}$ is a cyclic subgroup since it can be generated by the single element r . And since it has order n , then by Proposition 1.24, it must be isomorphic to $\mathbb{Z}_n$.

The quaternion group Q_8 (see Example 1.42) has the lattice diagram shown in Figure 2.6.

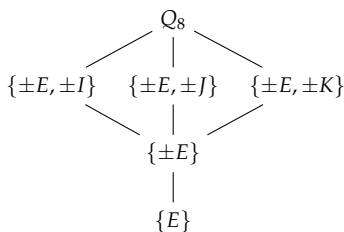


Figure 2.6: Lattice diagram for the quaternion group Q_8

This diagram is obviously different in structure to that for D_4 , and hence $D_4 \not\cong Q_8$.

Looking at the diagram for D_4 , we can see that

$$\begin{aligned}\{e, r^2\} &= \{e, r, r^2, r^3\} \cap \{e, r^2, m_2, m_4\} \\ &= \{e, r, r^2, r^3\} \cap \{e, r^2, m_1, m_3\} \\ &= \{e, r^2, m_2, m_4\} \cap \{e, r^2, m_1, m_3\}.\end{aligned}$$

Something similar happens in the other lattice diagrams we've seen so far: the intersection of any two subgroups is itself a subgroup, although in many cases this intersection will be either the trivial subgroup or one of the two intersecting subgroups. This is true in general:

Proposition 2.9 *Let $H, K < G$ be subgroups of some group G . Then their intersection $H \cap K$ is also a subgroup of G , and of H and K .*

Proof Since both H and K are subgroups of G , they both contain the identity e of G , and hence their intersection $H \cap K$ does as well, and is therefore not empty. Given elements $g_1, g_2 \in H \cap K$, it follows that both $g_1, g_2 \in H$ and $g_1, g_2 \in K$. Therefore their product $g_1 * g_2 \in H$ and also $g_1 * g_2 \in K$, hence $g_1 * g_2 \in H \cap K$. Thus $H \cap K$ is closed under the induced multiplication operation. Finally, we need to show that $H \cap K$ contains all the necessary inverses. But if $g \in H \cap K$ then $g \in H$ and so $g^{-1} \in H$; also $g \in K$ and so $g^{-1} \in K$ as well. Therefore $g^{-1} \in H \cap K$ as required, and hence $H \cap K$ is a subgroup of G .

Since $H \cap K \subseteq H$, and we have just seen that it forms a group under the induced multiplication operation, it must therefore be a subgroup of H . By an almost identical argument, $H \cap K$ is a subgroup of K . $\square$

We now introduce a related concept and proposition that will come in useful in the next section when we classify groups of order 6.

Definition 2.10 Let $H, K < G$ be two subgroups of a group G . Then denote by HK the set

$$HK = \{h * k : h \in H, k \in K\}.$$

If we are using additive notation, we may choose to denote this as

$$H + K = \{h + k : h \in H, k \in K\}.$$

Note that if G is nonabelian, it needn't be the case that $HK = KH$. If, however, the elements of H commute with the elements of K then $HK = KH$ is a subgroup of G :

Proposition 2.11 *Let H and K be subgroups of a group G . Then HK is a subgroup of G if and only if $HK = KH$.*

Proof For any $h \in H$ and $k \in K$ we have $(hk)^{-1} = k^{-1}h^{-1} \in KH$, so if HK is a subgroup of G then it must contain all its inverses, so $HK \subseteq KH$. Furthermore, $h = h * e \in HK$ and $k = e * k \in HK$. If HK is a

subgroup, it must be closed under multiplication, so $k*h$ must also lie in HK , and so $KH \subseteq HK$, which means that $HK = KH$ as required.

Conversely, suppose that $HK = KH$. To show that it is a subgroup of G , we need to confirm that it is closed under inverses and multiplication. For the first of these, suppose that $h \in H$ and $k \in K$, so that $hk \in HK$. Then $(hk)^{-1} = k^{-1}h^{-1} \in KH = HK$, so HK contains all required inverses.

Now suppose that $h_1, h_2 \in H$ and $k_1, k_2 \in K$. Then $(h_1k_1)(h_2k_2) = h_1(k_1h_2)k_2$. But $k_1h_2 \in KH$ and $KH = HK$, so there exists $h_3 \in H$ and $k_3 \in K$ such that $k_1h_2 = h_3k_3$, hence

$$(h_1k_1)(h_2k_2) = h_1(k_1h_2)k_2 = h_1(h_3k_3)k_2 = (h_1h_3)(k_3k_2) \in HK$$

and thus HK is closed under multiplication. It is therefore a subgroup of G , as claimed. $\square$

Proposition 2.12 Suppose $H, K < G$ are two subgroups of G such that $H \cap K = \{e\}$, every element of H commutes with every element of K (that is, $h * k = k * h$ for all $h \in H$ and $k \in K$), and that $HK = G$. Then $G \cong H \times K$.

This is sometimes called the **internal direct product** of H and K .

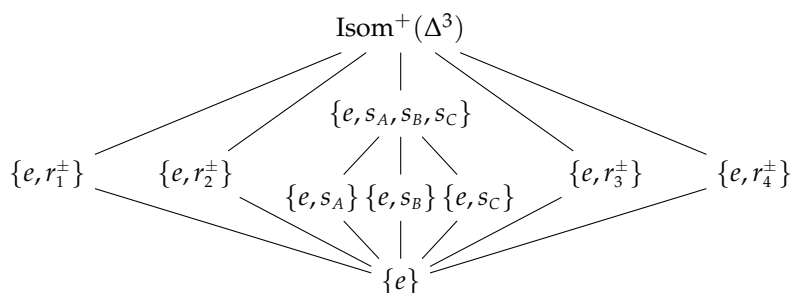
Proof Let $f: H \times K \rightarrow G$ be defined by $(h, k) \mapsto h * k$ for all $h \in H$ and $k \in K$. Then if $h_1, h_2 \in H$ and $k_1, k_2 \in K$, we have

$$\begin{aligned} f((h_1, k_1)(h_2, k_2)) &= f(h_1 * h_2, k_1 * k_2) = h_1 * h_2 * k_1 * k_2 \\ &= h_1 * k_1 * h_2 * k_2 = f(h_1, k_1) * f(h_2, k_2). \end{aligned}$$

This proves that f satisfies the structure condition for an isomorphism. It is clearly surjective, since we know that $G = HK$, and so any element of G can be expressed as a product $h * k = f(h, k)$, which is the image of some ordered pair $(h, k) \in H \times K$. All that remains is to show that f is injective. Suppose that $f(h_1, k_1) = f(h_2, k_2)$. Then $h_1 * k_1 = h_2 * k_2$ and so $h_2^{-1} * h_1 = k_2 * k_1^{-1}$. But the terms on the left-hand side of this expression belong to H , while the terms on the right-hand side belong to K . So both $h_2^{-1} * h_1$ and $k_2 * k_1^{-1}$ must belong to the intersection $H \cap K = \{e\}$, and therefore $h_2^{-1} * h_1 = k_2 * k_1^{-1} = e$, so $h_1 = h_2$ and $k_1 = k_2$, thus proving that f is injective, and hence the required isomorphism $H \times K \cong G$. $\square$

Another class of groups we met in the last chapter were the symmetric groups $\text{Sym}(X)$ of permutations on some set X , and in particular the finite symmetric groups S_n for $n \in \mathbb{N}$. Recall that we also studied the even permutations in S_n : those which can be expressed as an even number of transpositions. In the language of this chapter, we can certainly say that the alternating group A_n is a subgroup of S_n . But S_n has, in general, many other subgroups too: for example, as can be

readily seen from Figure 2.3, $S_3 \cong D_3$ has subgroups isomorphic to $\mathbb{Z}_2$ and $\mathbb{Z}_3$ as well as the trivial group and S_3 itself. As remarked in the last chapter, $\text{Isom}(\Delta^3) \cong S_4$ and $\text{Isom}^+(\Delta^3) \cong A_4$. By examining the multiplication table in Example 1.49, we can see that $\text{Isom}^+(\Delta^3)$ (and hence A_4 and S_4) has subgroups isomorphic to $\mathbb{Z}_2$ (such as $\{e, s_A\}$) and $\mathbb{Z}_3$ (such as $\{e, r_1^+, r_1^-\}$). It also contains a subgroup $\{e, s_A, s_B, s_C\}$ isomorphic to the Klein group. The lattice diagram is as follows:



The larger group $\text{Isom}(\Delta^3) \cong S_4$ contains an even wider collection of subgroups, including $\text{Isom}^+(\Delta^3)$ and all its subgroups, three subgroups of order 8, four of order 6, six more subgroups of order 4, and six more subgroups of order 2.

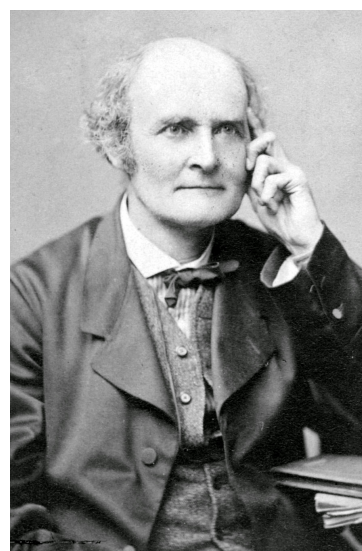
So far, we've found isomorphic copies of several small groups contained in symmetric groups, and it's tempting to ask whether this is going to be true in general: given an arbitrary group G , can we find a set X such that the symmetric group $\text{Sym}(X)$ has a subgroup which is isomorphic to G ? The answer is yes, as the following theorem demonstrates.

Theorem 2.13 (Cayley's Theorem) *Every group is isomorphic to a permutation group. That is, a subgroup of $\text{Sym}(X)$ for some (finite or infinite) set X .*

Proof For an element $g \in G$ we can define a function $\lambda_g: G \rightarrow G$ by $\lambda_g(h) = g * h$ for any $h \in G$. In the special case where g is the identity e , the function λ_e is just the identity map $\text{id}_G: G \rightarrow G$. More generally, λ_g is a bijection $G \rightarrow G$, and therefore a permutation of G . To show this, we need to check that it's both injective and surjective.

Injectivity follows from the left cancellation law in Proposition 1.15: if $\lambda_g(h) = \lambda_g(k)$ for two elements $h, k \in G$, then this means that $g * h = g * k$ and then (1.4) implies that $h = k$, so λ_g is injective.

Surjectivity can be shown as follows. Given some element $h \in G$, we want to find some element $k \in G$ such that $\lambda_g(k) = h$. But we can always do this by setting $k = g^{-1} * h$, since $\lambda_g(g^{-1} * h) = g * g^{-1} * h = h$ as required. So $\lambda_g: G \rightarrow G$ is bijective and thus a permutation of G .



Wikimedia Commons

The British mathematician **Arthur Cayley** (1821–1895) was one of the pioneers of group theory. At 17 he went to Trinity College, Cambridge, where he excelled: during his undergraduate career he published three papers in the *Cambridge Mathematical Journal*, and in 1842 came top in the examinations for the Mathematical Tripos, thereby winning the title of Senior Wrangler, the Smith's Prize and a four-year college fellowship.

Upon leaving Cambridge in 1846 he trained as a barrister at Lincoln's Inn and practiced law for the next fourteen years, during which he also wrote between two and three hundred mathematical papers, and formulated the Cayley–Hamilton Theorem in linear algebra (that a matrix satisfies its own characteristic equation). In 1863 he returned to Cambridge to become the first Sadleirian Professor of Pure Mathematics, in the process willingly giving up a well-paid legal career for a more modest academic salary, and welcoming the opportunity to devote more of his time to mathematics. During this phase of his career he continued his research on groups and the theory of invariants, much of it in collaboration with his near-contemporary James Sylvester (1814–1897). He also became a keen supporter of university education for women, and was involved in the foundation and early development of the University's two women's colleges, Girton and Newnham.

The key idea of the rest of this proof is to use these permutations λ_g to construct a subgroup of $\text{Sym}(G)$ which happens to be isomorphic to G itself.

Let $S = \{\lambda_g : g \in G\}$ be the set of these left-multiplication permutations. We now need to show that this subset $S \subseteq \text{Sym}(G)$ is actually a subgroup of $\text{Sym}(G)$, and furthermore that it's isomorphic to G .

The group operation on S is the induced operation from $\text{Sym}(G)$, which is just composition of permutations. So we need to show that S is closed under composition, and also that $\lambda_g^{-1} \in S$ for all $g \in G$.

To check closure we need to examine how composition works in S and how this relates to multiplication in G . Given $g, h, k \in G$ we have

$$(\lambda_g \circ \lambda_h)(k) = \lambda_g(\lambda_h(k)) = \lambda_g(h * k) = g * h * k = \lambda_{g*h}(k).$$

So $\lambda_g \circ \lambda_h = \lambda_{g*h}$, and since, for any $g, h \in G$, the product $g * h$ is also in G , it follows that λ_{g*h} is in S . Hence S is closed under composition.

Next we have to check the existence of inverses. As noted earlier, the function λ_e is the identity map id_G , which is just the identity permutation $\iota \in \text{Sym}(G)$. So the inverse λ_g^{-1} of λ_g is the permutation λ_h such that $\lambda_h \circ \lambda_g = \lambda_e = \lambda_g \circ \lambda_h$. But we know from the previous paragraph that $\lambda_g \circ \lambda_h = \lambda_{g*h}$, so what we're really looking for is $h \in G$ such that $h * g = e$, which rôle is clearly served by $h = g^{-1}$, and hence $\lambda_g^{-1} = \lambda_{g^{-1}}$.

So S is closed under composition, and contains both an identity element and a full set of inverses, and is therefore a subgroup of $\text{Sym}(G)$.

All that's left is to show $S \cong G$, which requires us to find a bijection $f: G \rightarrow S$ satisfying the structural condition in Definition 1.17.

An obvious candidate for this isomorphism f is the map that takes an element g to its corresponding permutation $\lambda_g \in S$; that is, $f: g \mapsto \lambda_g$.

Then f is injective, because if $f(g) = \lambda_g = \lambda_h = f(h)$ for some $g, h \in G$, it follows that $\lambda_g(k) = g * k = h * k = \lambda_h(k)$ for every element $k \in G$, and by the right cancellation law (1.5) it follows that $g = h$ and hence f is injective. The definition of S immediately confirms that f is surjective: $\lambda_g \in S$ precisely when $g \in G$.

All that remains is to show that f respects the group structure:

$$f(g) \circ f(h) = \lambda_g \circ \lambda_h = \lambda_{g*h} = f(g * h)$$

This completes the proof. □

To make sure we understand this, let's try a concrete example.

Example 2.14 Let $G = \mathbb{Z}_3$. Cayley's Theorem says we can find a permutation group isomorphic to $\mathbb{Z}_3 = \{0, 1, 2\}$, and by the construction in the above proof we know that we're actually looking for a

subgroup of $\text{Sym}(\mathbb{Z}_3) = S_3$.

The permutation λ_0 is obviously the identity permutation ι , since $\lambda_0(n) = 0 + n = n$.

The permutation λ_1 is defined by $\lambda_1(n) = 1 + n \pmod{3}$, which can be written as $\begin{bmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \end{bmatrix}$, or (012) in cycle notation.

The permutation λ_2 is defined by $\lambda_2(n) = 2 + n \pmod{3}$, which can be written as $\begin{bmatrix} 0 & 1 & 2 \\ 2 & 0 & 1 \end{bmatrix}$, or (021) in cycle notation.

So $\mathbb{Z}_3 \cong \{\lambda_0, \lambda_1, \lambda_2\} = \{\iota, (012), (021)\} < S_3$.

This example concerns a finite group, but there's nothing in our proof of Cayley's Theorem that relies on the groups G , $\text{Sym}(G)$ or S being finite, and in fact the theorem works for infinite groups too:

Example 2.15 Let $G = \mathbb{Z}$, which is countably infinite. Then for any integer $n \in \mathbb{Z}$, the function $\lambda_n: \mathbb{Z} \rightarrow \mathbb{Z}$ is defined by $\lambda_n(m) = n + m$ for all $m \in \mathbb{Z}$. This is clearly a bijection on $\mathbb{Z}$; it permutes the integers by shifting everything along by n places in the appropriate (positive or negative) direction. The group $S = \{\lambda_n : n \in \mathbb{Z}\}$ is a subgroup of the full permutation group $\text{Sym}(\mathbb{Z})$, and the isomorphism $f: \mathbb{Z} \rightarrow S$ that we want is precisely the function defined by mapping an integer n to the “shift everything along by n places” permutation λ_n .

Cayley's Theorem justifies the effort we spent studying permutations in Section 1.4: every group can be regarded as a permutation group.

2.2 Cosets and Lagrange's Theorem

IN THE PROOF OF CAYLEY'S THEOREM we introduced permutations $\lambda_g: G \rightarrow G$ which we used to show that any group can be viewed as a permutation group. These permutations are defined on the whole of G , but since part of our motivation for introducing the concept of a subgroup was to investigate the internal structure of groups, it's natural to ask how the λ_g behave when we restrict them to a subgroup.

The action on the trivial subgroup $\{e\}$ is obvious: $\lambda_g(e) = g * e = g$. To see what happens to a larger subgroup, let's look at the dihedral group D_3 and its rotation subgroup $R_3 = \{e, r, r^2\}$. We can read off the image $\lambda_g(R_3) = \{g, g * r, g * r^2\}$ from the multiplication table: it's just the part of the table where the row corresponding to the element g intersects with the columns corresponding to the elements of the subgroup R_3 . For example, $\lambda_{m_1}(R_3)$ is highlighted in Table 2.4. If we do this for all of D_3 we get

$$\lambda_e(R_3) = \{e, r, r^2\} = R_3, \quad \lambda_{m_1}(R_3) = \{m_1, m_2, m_3\} = D_3 \setminus R_3,$$

The reader will find no figures in this work. The methods which I set forth do not require either constructions or geometrical or mechanical reasonings: but only algebraic operations, subject to a regular and uniform rule of procedure.

— Joseph-Louis Lagrange (1736–1813),
preface to *Mécanique Analytique* (1788)

$\cdot$	e	r	r^2	m_1	m_2	m_3
e	e	r	r^2	m_1	m_2	m_3
r	r	r^2	e	m_3	m_1	m_2
r^2	r^2	e	r	m_2	m_3	m_1
m_1	m_1	m_2	m_3	e	r	r^2
m_2	m_2	m_3	m_1	r^2	e	r
m_3	m_3	m_1	m_2	r	r^2	e

Table 2.4: The multiplication table for the dihedral group D_3 with the image $\lambda_{m_1}(R_3) = \{m_1, m_2, m_3\}$ highlighted

$$\begin{aligned}\lambda_r(R_3) &= \{r, r^2, e\} = R_3, & \lambda_{m_2}(R_3) &= \{m_2, m_3, m_1\} = D_3 \setminus R_3, \\ \lambda_{r^2}(R_3) &= \{r^2, e, r\} = R_3, & \lambda_{m_3}(R_3) &= \{m_3, m_1, m_2\} = D_3 \setminus R_3.\end{aligned}$$

Immediately, we notice that (apart from some inconsequential reordering) the image $\lambda_g(R_3)$ has one of two different forms. If $g \in R_3$ then $\lambda_g(R_3) = R_3$. (This is exactly what we'd expect, since R_3 , being a subgroup, is closed under the induced multiplication operation.) If, on the other hand, $g \notin R_3$ then $\lambda_g(R_3) = \{m_1, m_2, m_3\} = D_3 \setminus R_3$.

So the action of the various permutations on the subgroup R_3 neatly partitions D_3 into two different subsets. These images $\lambda_g(R_3)$ look like they might be a useful concept, so we give them a special name:

Definition 2.16 Let G be a group, and suppose $H < G$ is a subgroup of G . Then, for a given element $g \in G$, the **left coset** gH is defined to be the subset $\lambda_g(H) = \{g * h : h \in H\} \subseteq G$.

If we happen to be using additive notation for our group G , then we will usually denote left cosets as $g+H$ rather than gH .

Recall that we could also have proved Cayley's Theorem by means of the permutations $\rho_g: G \rightarrow G$ defined by $\rho_g(h) = h * g$. In the interests of completeness and achirality, we introduce the following definition:

Definition 2.17 Let G be a group, and suppose $H < G$ is a subgroup of G . Then, for a given element $g \in G$, the **right coset** Hg is defined to be the subset $\rho_g(H) = \{h * g : h \in H\} \subseteq G$.

$\cdot$	e	r	r^2	m_1	m_2	m_3
e	e	r	r^2	m_1	m_2	m_3
r	r	r^2	e	m_3	m_1	m_2
r^2	r^2	e	r	m_2	m_3	m_1
m_1	m_1	m_2	m_3	e	r	r^2
m_2	m_2	m_3	m_1	r^2	e	r
m_3	m_3	m_1	m_2	r	r^2	e

Table 2.5: The multiplication table for the dihedral group D_3 with the image $\rho_{m_1}(R_3) = \{m_1, m_2, m_3\}$ highlighted

Again, if we're using additive notation for G , then we will typically denote right cosets as $H+g$ rather than Hg .

We can easily read off right cosets from the multiplication table too. Just as the left cosets corresponded to parts of a particular row of the table, right cosets correspond to parts of a particular column (see Table 2.5).

The right cosets of R_3 in D_3 are

$$\begin{aligned}R_3e &= \rho_e(R_3) = \{e, r, r^2\} = R_3, & R_3m_1 &= \rho_{m_1}(R_3) = \{m_1, m_3, m_2\} = D_3 \setminus R_3, \\ R_3r &= \rho_r(R_3) = \{r, r^2, e\} = R_3, & R_3m_2 &= \rho_{m_2}(R_3) = \{m_2, m_1, m_3\} = D_3 \setminus R_3, \\ R_3r^2 &= \rho_{r^2}(R_3) = \{r^2, e, r\} = R_3, & R_3m_3 &= \rho_{m_3}(R_3) = \{m_3, m_2, m_1\} = D_3 \setminus R_3.\end{aligned}$$

Comparing this list of right cosets with the left cosets on page 49, we find that the two lists are the same. Or, more formally,

$$R_3g = \rho_g(R_3) = \lambda_g(R_3) = gR_3$$

for all $g \in D_3$. We might be tempted to conjecture that this is true in general, and it clearly is in the case where G is abelian, since

$$\rho_g(h) = h * g = g * h = \lambda_g(h)$$

for all $g \in G$ and $h \in H < G$.

This isn't always true if G is nonabelian, as the next example shows.

Example 2.18 Let $M = \{e, m_1\} < D_3$. Then

$$\begin{aligned} eM &= m_1M = M, & Me &= Mm_1 = M, \\ rM &= m_3M = \{r, m_3\}, & Mr &= Mm_2 = \{r, m_2\}, \\ r^2M &= m_2M = \{r^2, m_2\}, & Mr^2 &= Mm_3 = \{r^2, m_3\}. \end{aligned}$$

So, in particular,

$$rM = m_3M \neq Mm_3 = Mr^2 \quad \text{and} \quad r^2M = m_2M \neq Mm_2 = Mr.$$

It's interesting to ask what special underlying property R_3 has, and M doesn't, that causes its left and right cosets to line up in this way. We'll leave that question until the next chapter, where it will motivate our study of normal subgroups and quotient groups. For the moment, left and right cosets will keep us busy enough.

Let's look at a few more examples. We'll start with a couple of small abelian groups we've met several times already.

Example 2.19 The Klein group $V_4 = \{e, a, b, c\}$ has three proper, nontrivial subgroups

$$A = \{e, a\}, \quad B = \{e, b\}, \quad C = \{e, c\}.$$

The group V_4 is abelian, so the left and right cosets coincide, and are as follows:

$$\begin{aligned} eA &= aA = \{e, a\} = Ae = Aa, & bA &= cA = \{b, c\} = Ab = Ac, \\ eB &= bB = \{e, b\} = Be = Bb, & aB &= cB = \{a, c\} = Ba = Bc, \\ eC &= cC = \{e, c\} = Ce = Cc, & aC &= bC = \{a, b\} = Ca = Cb. \end{aligned}$$

	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

Table 2.6: Multiplication table for the Klein group V_4

Example 2.20 Consider the cosets of $H = \{0, 3\} < \mathbb{Z}_6$. This is an abelian group, so the left and right cosets will coincide.

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

The subgroup $K = \{0, 2, 4\}$ has cosets

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

The full group $\mathbb{Z}_6$ has just one coset: itself, since $n+\mathbb{Z}_6 = \mathbb{Z}_6$ for all $n \in \mathbb{Z}_6$. Finally, the trivial subgroup $\{0\}$ has six cosets, since $n+\{0\} = \{n\}$ for all $n \in \mathbb{Z}_6$.

	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

Table 2.7: Multiplication table for the cyclic group $\mathbb{Z}_6$

Now let's look at a nonabelian group and see what happens.

Example 2.21 The quaternion group Q_8 (see Table 2.8) has four proper, nontrivial subgroups, three of which are isomorphic to $\mathbb{Z}_4$ and one which is isomorphic to $\mathbb{Z}_2$:

$$A = \{\pm E, \pm I\}, \quad B = \{\pm E, \pm J\}, \quad C = \{\pm E, \pm K\}, \quad D = \{\pm E\}.$$

	E	I	J	K	$-E$	$-I$	$-J$	$-K$
E	E	I	J	K	$-E$	$-I$	$-J$	$-K$
I	I	$-E$	K	$-J$	$-I$	E	$-K$	J
J	J	$-K$	$-E$	I	$-J$	K	E	$-I$
K	K	J	$-I$	$-E$	$-K$	$-J$	I	E
$-E$	$-E$	$-I$	$-J$	$-K$	E	I	J	K
$-I$	$-I$	E	$-K$	J	I	$-E$	K	$-J$
$-J$	$-J$	K	E	$-I$	J	$-K$	$-E$	I
$-K$	$-K$	$-J$	I	E	K	J	$-I$	$-E$

Table 2.8: Multiplication table for the quaternion group Q_8

The cosets of A are as follows:

$$\begin{aligned}(E)A &= (-E)A = (I)A = (-I)A = \{\pm E, \pm I\}, \\ (J)A &= (-J)A = (K)A = (-K)A = \{\pm J, \pm K\}, \\ A(E) &= A(-E) = A(I) = A(-I) = \{\pm E, \pm I\}, \\ A(J) &= A(-J) = A(K) = A(-K) = \{\pm J, \pm K\}.\end{aligned}$$

The cosets of B and C follow a similar pattern (check this). We see from the above list (and the corresponding ones for B and C) that the left and right cosets of these three subgroups do coincide.

Looking at the cosets of $\{\pm E\}$ we find that

$$\begin{aligned}(E)D &= (-E)D = \{\pm E\}, & (I)D &= (-I)D = \{\pm I\}, \\ (J)D &= (-J)D = \{\pm J\}, & (K)D &= (-K)D = \{\pm K\}, \\ D(E) &= D(-E) = \{\pm E\}, & D(I) &= D(-I) = \{\pm I\}, \\ D(J) &= D(-J) = \{\pm J\}, & D(K) &= D(-K) = \{\pm K\}.\end{aligned}$$

In this case as well, the corresponding left and right cosets coincide.

All the examples we've seen so far are of finite groups. These concepts carry across to the case where the group under investigation is infinite.

Example 2.22 Consider the (infinite) cyclic subgroup $3\mathbb{Z} = \langle 3 \rangle$ of the (infinite) cyclic group $\mathbb{Z}$. This has three distinct left cosets (which, since $\mathbb{Z}$ is abelian, are each identical to the corresponding right coset):

$$\begin{aligned}0+3\mathbb{Z} &= \{\dots, -6, -3, 0, 3, 6, \dots\} = 3\mathbb{Z}+0, \\ 1+3\mathbb{Z} &= \{\dots, -5, -2, 1, 4, 7, \dots\} = 3\mathbb{Z}+1, \\ 2+3\mathbb{Z} &= \{\dots, -4, -1, 2, 5, 8, \dots\} = 3\mathbb{Z}+2.\end{aligned}$$

Example 2.23 Recall that the orthogonal group $O_3(\mathbb{R})$ consists of all 3×3 real orthogonal matrices; that is, matrices A such that $A^{-1} = A^T$. These must have determinant ± 1 since if $A^T = A^{-1}$ then $\det A^T = \det(A^{-1}) = (\det A)^{-1}$. But also $\det A^T = \det A$ in general. By putting these together we have $\det A = (\det A)^{-1}$, or $(\det A)^2 = 1$ which implies that $\det A = \pm 1$.

The special orthogonal group $SO_3(\mathbb{R})$ is the subgroup of orthogonal matrices which have determinant $+1$. It has two cosets, namely $SO_3(\mathbb{R})$ itself and the complement $O_3(\mathbb{R}) \setminus SO_3(\mathbb{R})$, comprising all orthogonal matrices with determinant -1 .

For any special orthogonal matrix A , we see that

$$A SO_3(\mathbb{R}) = SO_3(\mathbb{R}) = SO_3(\mathbb{R}) A$$

since $SO_3(\mathbb{R})$ is closed under matrix multiplication.

If, on the other hand, B is orthogonal with determinant -1 , then BA is also orthogonal, because $O_3(\mathbb{R})$ is itself a group and hence closed

under matrix multiplication. The product BA also has determinant -1 since $\det(BA) = (\det B)(\det A) = -1 \times 1 = -1$. So

$$B SO_3(\mathbb{R}) = O_3(\mathbb{R}) \setminus SO_3(\mathbb{R}) = SO_3(\mathbb{R}) B.$$

Example 2.24 In Example 2.5 we met the circle subgroup $C < \mathbb{C}_*^\times$ of the multiplicative group of nonzero complex numbers. Now let's look at its cosets. Complex multiplication is commutative, so we know straight away that the left and right cosets will coincide.

An element of C can be written in the form $e^{i\theta}$ where $0 \leq \theta < 2\pi$; any other nonzero complex number can be written in polar form as $re^{i\phi}$ where $r > 0$ and $0 \leq \phi < 2\pi$.

Multiplying these together, we get $re^{i(\theta+\phi)}$, which can be regarded geometrically as a point on a circle of radius r centred on the origin. The subgroup C is obviously closed, so $zC = C$ for any $z \in C$ where $|z| = 1$. If $|z| = r \neq 1$ then the coset zC will be the circle of radius r centred on the origin. Thus C has an uncountably infinite number of cosets in $\mathbb{C}_*^\times$, each with an uncountably infinite number of elements. These are shown in Figure 2.7

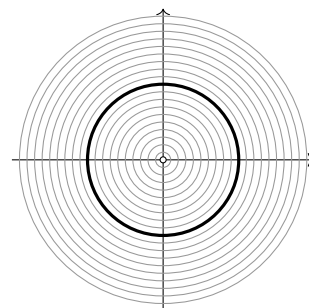


Figure 2.7: Cosets of $C < \mathbb{C}_*^\times$

Example 2.25 The multiplicative group $\mathbb{R}_*^\times$ of nonzero real numbers can also be regarded as a subgroup of $\mathbb{C}_*^\times$. For $z \in \mathbb{R}_*^\times < \mathbb{C}_*^\times$, the coset $z\mathbb{R}_*^\times$ is just $\mathbb{R}_*^\times$ itself. If a nonzero complex number $z \notin \mathbb{R}_*^\times$ then the coset $z\mathbb{R}_*^\times = \{zx : x \in \mathbb{R}_*^\times\}$, the set of all nonzero real multiples of the nonzero complex number z . Geometrically, this is the straight line through the origin, passing through the point z , with the origin itself deleted. Again, we have an uncountably infinite number of cosets, each containing an uncountably infinite number of elements. These cosets are shown in Figure 2.8

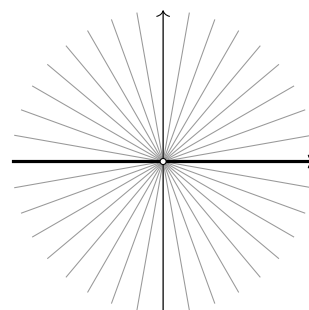


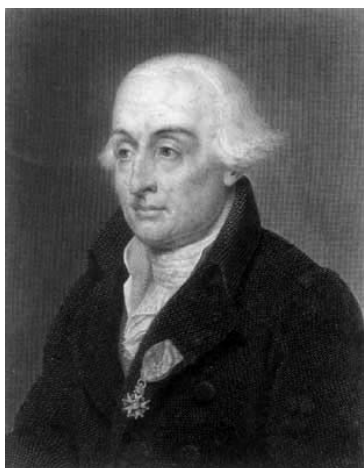
Figure 2.8: Cosets of $\mathbb{R}_*^\times < \mathbb{C}_*^\times$

One thing all of these examples have in common is that the cosets **partition** the group: their union is the whole group, and no two cosets intersect, so the group neatly splits into a (possibly infinite) collection of non-overlapping cosets. The next proposition confirms this is true in general; more importantly it will help us prove Lagrange's Theorem, an important result about the internal structure of finite groups.

Proposition 2.26 *Let $H < G$ be a subgroup of a (possibly infinite) group G , and let $g, k \in G$ be two arbitrary elements of G . Then the following three statements are equivalent:*

- (i) $k \in Hg$,
- (ii) $Hg = Hk$, and
- (iii) $kg^{-1} \in H$.

(There is a corresponding result for left cosets.)



Wikimedia Commons / Robert Hart (fl.1830s)

Joseph-Louis Lagrange (1736–1813), born Giuseppe Lodovico Lagrangia in Turin, was an Italian and French mathematician and physicist who made many important contributions to algebra, analysis, number theory and classical mechanics. His *Mécanique Analytique* (1788) was the most comprehensive treatment of classical mechanics since Newton's celebrated *Philosophiæ Naturalis Principia Mathematica* (1687).

In 1766 he succeeded Leonhard Euler as director of mathematics at the Prussian Academy of Sciences in Berlin and remained there for twenty years, moving to Paris in 1786 at the invitation of Louis XVI. His first few years in Paris were marked by a prolonged episode of depression and the political instability of the Revolution; he survived both and was appointed to a chair in analysis at the École Polytechnique in 1794, and a chair in mathematics at the École Normale in 1795. Contemporary accounts indicate that he was considerably less gifted at teaching than at research.

He was elected a Fellow of the Royal Society of London in 1806, appointed a *Grand Officier* of the *Légion d'Honneur*, and a *Comte de l'Empire* by Napoleon in 1808, who later awarded him the *Grand Croix* of the *Ordre Impérial de la Réunion* a week before his death in 1813. He was granted the honour of a tomb in the crypt of the Panthéon, and he is one of 72 eminent French scientists commemorated on the Eiffel Tower.

Proof First we prove that (i) $\implies$ (ii). Suppose that $k \in Hg$. Then there exists a unique $h \in H$ such that $h * g = k$. Multiplying on the left by h^{-1} gives $g = h^{-1} * k$. As usual, to prove $Hg = Hk$ we need to show that $Hg \subseteq Hk$ and $Hk \subseteq Hg$. So, choose some element $a \in Hg$. Then $a = b * g$ for some $b \in H$, and so $a = b * h^{-1} * k$, which means that $a \in Hk$ (since $b * h^{-1} \in H$ as both $b, h^{-1} \in H$). Hence $Hg \subseteq Hk$. Similarly, choose $c \in Hk$. Then $c = d * k$ for some $d \in H$, so $c = d * h * g$, which means that $c \in Hg$ (because $d * h \in H$ as both $d, h \in H$). Hence $Hk \subseteq Hg$ and therefore $Hg = Hk$.

The converse (ii) $\implies$ (i) holds as well: if $Hg = Hk$ then since $k \in Hk$, it follows that $k \in Hg$ as well.

Next we show that (i) $\implies$ (iii). If $k \in Hg$ then, as noted above, there is a unique $h \in H$ such that $h * g = k$. Multiplying on the right by g^{-1} gives $k * g^{-1} = h \in H$ as required.

Finally, we show the converse (iii) $\implies$ (i). If $h = k * g^{-1} \in H$ then it follows almost immediately that $k = hg \in Hg$. $\square$

This proposition has a couple of important corollaries:

Corollary 2.27 Two right cosets Hg and Hk are either equal or disjoint.

Proof Suppose $a \in Hg \cap Hk$. Then, by Proposition 2.26, since $a \in Hg$, it follows that $Ha = Hg$; also since $a \in Hk$ we have $Ha = Hk$ and hence $Hg = Hk$. The alternative is that no such element a exists, in which case $Hg \cap Hk = \emptyset$. $\square$

Corollary 2.28 The right cosets of H in G partition G .

The next proposition tells us an important fact about the relative sizes of right cosets of a finite subgroup $H < G$.

Proposition 2.29 Let H be a subgroup of a (possibly infinite) group G . Then $|Hg| = |H|$ for any element $g \in G$.

Proof By the right cancellation law in Proposition 1.15, if $h_1 * g = h_2 * g$ for some $h_1, h_2 \in H$ and $g \in G$, then $h_1 = h_2$. This means that the function $f: H \rightarrow Hg$ given by $h \mapsto h * g$ is a bijection, so $|H| = |Hg|$ as required. $\square$

These last two facts, Corollary 2.28 and Proposition 2.29, give the following elegant result.

Theorem 2.30 (Lagrange's Theorem) Let H be a subgroup of a finite group G . Then the order $|H|$ of H divides the order $|G|$ of G .

We can use this fact, that the order of a subgroup H is a factor of the whole group G , to devise some measure of the relative sizes of G and H , simply by looking at the quotient $|G|/|H|$. Lagrange's Theorem tells us that this will always be a positive integer.

For example, the rotation subgroup $R_3 < D_3$ is clearly half the size of D_3 itself, so this quotient $|D_3|/|R_3| = 2$. Similarly, $|S_n|/|A_n| = 2$ for any $n \in \mathbb{N}$. The subgroup $\langle 6 \rangle < \mathbb{Z}_{12}$ is obviously smaller relative to the whole group $\mathbb{Z}_{12}$, and indeed the quotient $|\mathbb{Z}_{12}|/|\langle 6 \rangle| = 6$.

So, the larger this quotient is, the smaller the subgroup in question is relative to the full group. We might wonder if this concept can be extended to the case of infinite groups, and indeed it can. The quotient $|G|/|H|$ is really just the number of distinct right cosets of H in G , so by looking at this way we can formulate the following definition:

Definition 2.31 The **index** $|G:H|$ of a subgroup $H < G$ is the number of right cosets of H in G . If G is finite then $|G:H| = |G|/|H|$.

The index of a subgroup is, in some sense, multiplicative:

Proposition 2.32 Suppose H and K are subgroups of some finite group G , such that $K < H < G$. Then $|G : K| = |G : H| \times |H : K|$.

Proof Since G is finite (and hence so are H and K) we have

$$|G : K| = |G|/|K| = |G|/|H| \cdot |H|/|K| = |G : H| \cdot |H : K|$$

as required. $\square$

We can extend this result to the case where G , H and K may be infinite: as long as $|G : H|$, $|G : K|$ and $|K : H|$ are all finite, the multiplicativity result still holds, although the proof is more involved and is left as an exercise to the reader.

The following is an easy application of Lagrange's Theorem.

Corollary 2.33 A group G of prime order $p = |G|$ has no proper non-trivial subgroups.

Proof The order of any subgroup $H < G$ must divide $p = |G|$ by Lagrange's Theorem. The only possibilities are $|H| = 1$ (in which case H is the trivial subgroup) or $|H| = p$ (in which case $H = G$). $\square$

With a little bit more effort, we can go further than this and demonstrate that there is only one group (up to isomorphism) of a given prime order p . First we need the following simple proposition, which relates the order of an element to the order of the group.

Proposition 2.34 Let G be a finite group. Then the order $|g|$ of any element $g \in G$ divides the order $|G|$ of the whole group G .

Proof Suppose $|g| = n$. Then the powers $\{g^i : i \in \mathbb{Z}\}$ form a finite subgroup of G ; this is the cyclic subgroup $\langle g \rangle$ generated by g . But $\langle g \rangle = \{g^i : 0 \leq i < n\}$ and so $|\langle g \rangle| = |g| = n$. So, by Lagrange's Theorem, $|\langle g \rangle| = |g| = n$ must be a factor of $|G|$. $\square$

This enables us to prove the following extension of Corollary 2.33:

Proposition 2.35 *Let G be a finite group of prime order $p = |G|$. Then $G \cong \mathbb{Z}_p$, the finite cyclic group of order p .*

Proof Suppose $g \in G$ is some element of G apart from the identity e . Then $|g| > 1$ by Proposition 1.19. By Proposition 2.34 above, $|g|$ must be a factor of $|G| = p$; since p is prime and we already know $|g| \neq 1$, then $|g| = p$. The cyclic subgroup $\langle g \rangle$ generated by g must therefore have p elements, and be isomorphic to $\mathbb{Z}_p$. It must also be the full group, hence $G \cong \mathbb{Z}_p$. $\square$

This is a particularly useful result from the point of view of compiling a list of finite groups of small order, since it tells us exactly what groups there are of a given prime order p : there's just one (up to isomorphism) and it's $\mathbb{Z}_p$. In a little while we'll develop further techniques to classify all groups of order 8 or less.

Lagrange's Theorem gives a very useful necessary condition on the order of subgroups of a finite group. But we must be careful not to read into it more than it actually says. For a start, it only applies to finite groups, not infinite ones. Also, the condition in Lagrange's Theorem is necessary but not sufficient: it says that the order of any subgroup must be a factor of the order of the whole group, but it doesn't guarantee that there *will* be a subgroup with order equal to a given factor. The smallest counterexample is the alternating group A_4 :

Example 2.36 By Lagrange's Theorem, all subgroups of the alternating group A_4 must have order equal to a factor $|A_4| = 12$; that is, if $H < A_4$ then $|H| = 1, 2, 3, 4, 6$ or 12 .

The trivial group $\{\iota\}$ consisting of the identity permutation ι is, of course, the only subgroup of order 1.

There are three subgroups of order 2:

$$\{\iota, (12)(34)\}, \quad \{\iota, (13)(24)\}, \quad \{\iota, (14)(23)\},$$

each consisting of the identity and an element of order 2. (Individual transpositions such as (12) or (23) have order 2, but they're odd permutations and hence not elements of A_4 .)

There are four subgroups of order 3, each generated by a 3-cycle:

$$\begin{aligned} \{\iota, (123), (132)\}, & \quad \{\iota, (124), (142)\}, \\ \{\iota, (134), (143)\}, & \quad \{\iota, (234), (243)\}. \end{aligned}$$

There is a single subgroup of order 4:

$$\{\iota, (12)(34), (13)(24), (14)(23)\}$$

(This happens to be isomorphic to the Klein 4-group V_4 ; in a short while we'll see that it couldn't be anything else.)

There are no other order-4 subgroups, because by Proposition 2.34, such a subgroup must comprise elements of order 1, 2 or 4, and the only remaining elements of A_4 are the eight 3-cycles, which have order 3.

So far, we've found subgroups of order 1, 2, 3, 4 and (A_4 itself) 12. The only remaining possibility, according to Lagrange's Theorem, is a subgroup of order 6. Suppose $H < A_4$ is such a subgroup. If a given 3-cycle belongs to H then its inverse (also a 3-cycle) must as well, so the number of 3-cycles in H is even. There can't be six of them, since that would leave no room for the identity permutation ι . Suppose that H contains four 3-cycles: α , α^{-1} , β and β^{-1} . Then $\iota, \alpha, \alpha^{-1}, \beta, \beta^{-1}, \alpha\beta$ and $\alpha\beta^{-1}$ are all distinct elements of H . But there are seven of them, which contradicts our hypothesis that H has six elements.

If H contains only two 3-cycles, then the rest of H must consist of the remaining four elements of A_4 , namely $\iota, (12)(34), (13)(24)$ and $(14)(23)$. But we noted earlier that these four elements form a subgroup of order 4. Lagrange's Theorem tells us that any subgroup of A_4 contained in H (and which must therefore itself be a subgroup of H) must have order a factor of $|H| = 6$. But clearly 4 isn't a factor of 6, so this 4-element group can't be a subgroup of H .

At this point we have run out of possibilities: if H contains no 3-cycles at all, then there aren't enough other elements in A_4 to make up a group of order 6, so our hypothesised subgroup H can't exist, and hence A_4 has no subgroups of order 6.

This example demonstrates an important general fact in mathematics: just because an object satisfying certain properties *might* exist, it doesn't mean that it actually *does*.

So, the converse of Lagrange's Theorem isn't true in general, which might seem like a bit of a setback to our attempts to understand the internal structure of groups. But all is not lost: while the full converse doesn't hold in all cases, there are a number of other results which provide a partial converse in certain cases. The Norwegian mathematician Ludwig Sylow's three celebrated theorems, for example, are a little sophisticated for this stage in the narrative, and we'll postpone further discussion until Chapter 7, but the following useful theorem is within our present means.

Theorem 2.37 (Cauchy's Theorem) *If p is a prime factor of the order of a finite group G , then G contains a nontrivial element (and hence a cyclic subgroup) of order p .*

This result, due to the French mathematician Augustin-Louis Cauchy

(1789–1857), says that although subgroups of arbitrary factors might not exist, cyclic subgroups of prime factors do. In Example 2.36, we saw that while A_4 doesn't have a subgroup of order 6, it does have subgroups of order 2 and 3, which are prime factors of $|A_4| = 12$.

Proof Let

$$X = \{(g_1, \dots, g_p) : g_1, \dots, g_p \in G \text{ and } g_1 * \dots * g_p = e\}$$

be the set of all ordered p -tuples of elements of G whose product is the identity e . What we're looking for is such a p -tuple where $g_1 = \dots = g_p \neq e$; that is, a p -tuple of the form $(g, \dots, g)$ which isn't $(e, \dots, e)$ but which does multiply out to give the identity.

We want, first of all, to know how big X is; in fact we claim that $|X|$ is a multiple of p . For any p -tuple $(g_1, \dots, g_p) \in X$ we know that since $g_1 * \dots * g_p = e$ it follows that $(g_1 * \dots * g_{p-1}) = g_p^{-1}$. So we effectively have a free choice for all of the first $(p-1)$ elements $g_1, \dots, g_{p-1}$ and then we just have to set $g_p = (g_1 * \dots * g_{p-1})^{-1}$ to ensure that $(g_1, \dots, g_p) \in X$. So, to write down all the p -tuples in X , we have $|G|$ choices for g_1 , $|G|$ choices for g_2 , and so on up to g_{p-1} . Hence $|X| = |G|^{p-1}$. But we also know from the hypothesis that p is a prime factor of $|G|$, so it must also be a factor of $|X| = |G|^{p-1}$; in other words, $|X|$ is, as claimed, a multiple of p .

Now, given a p -tuple $(g_1, \dots, g_p) \in X$, we can obtain another (possibly different) one by **cyclically permuting** the terms in the p -tuple:

$$(g_1, \dots, g_p) \mapsto (g_2, \dots, g_p, g_1).$$

This is still an element of X , since if $g_1 * \dots * g_p = e$ then

$$g_2 * \dots * g_p * g_1 = g_1^{-1} * (g_1 * \dots * g_p) * g_1 = g_1^{-1} * e * g_1 = g_1^{-1} * g_1 = e.$$

Now define an equivalence relation $\sim$ on the set X as follows. Given two p -tuples $\mathbf{x}, \mathbf{y} \in X$, we say that $\mathbf{x} \sim \mathbf{y}$ if $\mathbf{y}$ can be obtained from $\mathbf{x}$ by a finite number of cyclic permutations. For example,

$$(g_1, \dots, g_p) \sim (g_4, g_5, \dots, g_p, g_1, g_2, g_3)$$

because the second p -tuple can be obtained by cycling the first one 3 times.

This is an equivalence relation (check this yourself to make sure) and hence partitions X into a finite number of disjoint subsets: the equivalence classes of the relation $\sim$.

The p -tuple $(e, \dots, e)$ is unchanged by cyclic (or indeed any) permutation, so it forms a one-element equivalence class on its own.

The equivalence classes partition X , so if we sum their respective sizes we must get the number of elements in X . We know that at least one of the other equivalence classes must have fewer than p elements:

What's actually going on here is that we're defining an **action** of the symmetric group S_p on the set X : for each permutation $\sigma \in S_p$ we map

$$(g_1, \dots, g_p) \mapsto (g_{\sigma(1)}, \dots, g_{\sigma(p)}).$$

The cyclic permutation is obtained by using the full p -cycle $(12 \dots p) \in S_p$. We'll come back to these ideas in much more detail in Chapter 6.

none of them can have more than p elements since there are at most p distinct cyclic permutations of a given p -tuple, and if all of them had p elements, the total number of p -tuples in X would be of the form $mp+1$ for some integer m ; we know that it can't because we showed a little earlier that $|X|$ has to be a multiple of p .

So there must be at least one equivalence class with fewer than p elements.

If this class consists of just one p -tuple then we've found the order- p element we're looking for: any such p -tuple must have $g_1 = \cdots g_p$ otherwise cyclic permutation would yield at least one other distinct p -tuple; given such an element $(g, \dots, g)$ we immediately have $g^p = e$ as required.

So, suppose we have an equivalence class comprising fewer than p but more than one distinct p -tuples. Then two of them must be equal:

$$(g_{r+1}, \dots, g_p, g_1, \dots, g_r) = (g_{s+1}, \dots, g_p, g_1, \dots, g_s)$$

Suppose, without loss of generality, that $r < s$ and then cycle back r times (or forward $(p-r)$ times) to get

$$(g_1, \dots, g_p) = (g_{k+1}, \dots, g_p, g_1, \dots, g_k)$$

where $k = s - r$. Then we find that $g_i = g_{k+i}$ for all $1 \leq i \leq p$ and so

$$g_1 = g_{k+1} = g_{2k+1} \pmod{p} = \cdots = g_{(p-1)k+1} \pmod{p}. \quad (2.1)$$

Now suppose $ak + 1 \equiv bk + 1 \pmod{p}$ where $0 \leq a < b \leq p-1$. Then p divides $(b-a)k$. But this can't happen, because p is prime and both $(b-a)$ and k are strictly less than p .

So the numbers $1, k+1, 2k+1, \dots, (p-1)k+1$ are all different modulo p . Furthermore, there are p of them, and so (modulo p) we have some permutation of the numbers $1, \dots, p$. Substituting these into (2.1) and rearranging, we get

$$g_1 = g_2 = \cdots = g_p.$$

Hence our p -tuple $(g_1, \dots, g_p) = (g_1, \dots, g_1)$, and multiplying everything out we get $g_1^p = e$ as required. $\square$

We can use this theorem to help classify (that is, make a complete list of) groups of order 4, 6 and 8. (Groups of orders 1, 2, 3, 5 and 7 have already been classified: the trivial group is the only possible group of order 1, and by Proposition 2.35 there is only a single group, up to isomorphism, of a given prime order p , namely $\mathbb{Z}_p$.)

We start with groups of order 4:

Proposition 2.38 *A group of order 4 must be isomorphic either to the cyclic group $\mathbb{Z}_4$ or the Klein 4-group V_4 .*



Wikimedia Commons / Zéphirin Belliard (1798–1861)
after Jean Roller (1798–1866)

Augustin-Louis Cauchy (1789–1857) made numerous contributions to many areas of mathematics: during his career he published 789 papers and several books on mechanics, real and complex analysis, number theory, algebra and geometry, and has more concepts and theorems named after him than any other mathematician.

After a short career as a civil engineer in Cherbourg, he returned to Paris in 1812 and undertook research in mathematics, publishing papers on geometry and algebra. Over the next few years he was appointed to chairs at the École Polytechnique, the Collège de France and the University of Paris, and elected to the Académie des Sciences.

Following the July Revolution in 1830 he refused to swear allegiance to the new king, Louis Philippe I, and was dismissed from his academic posts. He subsequently became tutor to the exiled Henri d'Artois, nephew and heir to the previous king Charles X. This was a disaster: Henri acquired a lifelong hatred of mathematics and Cauchy was unable to do much research, although he was granted the title of Baron in recognition of his service.

Returning to Paris in 1838, he continued his research but didn't regain his academic posts until Louis Philippe was deposed ten years later. He remained a professor at the University until his death at the age of 67. He is one of 72 eminent scientists whose names are inscribed on the Eiffel Tower.

Proof Let $G = \{e, a, b, c\}$ be a group with 4 elements. Proposition 2.34 tells us that the order of each of these elements must be a factor of $|G| = 4$. The identity element e has order 1, and by Proposition 1.19 is the only element which does. So the remaining elements must have order either 2 or 4. Cauchy's Theorem asserts the existence of at least one element of order 2. If at least one of the other elements has order 4, say a , then that means G has a cyclic subgroup $\{e, a, a^2, a^3\} \cong \mathbb{Z}_4$, which accounts for all of the elements in G and hence $G \cong \mathbb{Z}_4$. Here a^2 is the element of order 2 whose existence was guaranteed by Cauchy's Theorem.

If a , b and c all have order 2 then the cancellation laws show that the product of any two must be the third. For example, $a * b = b$ implies that $a = e$, which can't be true since we assumed a has order 2; similarly $a * b = a$ forces $b = e$ which also can't be true for the same reason. Neither can $a * b = e$, because that would imply $a = b^{-1}$; this can't be true since a and b both have order 2 and so $a = a^{-1}$ and $b = b^{-1}$. The only remaining possibility is that $a * b = c$. This is the Klein group V_4 . $\square$

A similar argument enables us to classify groups of order 6:

Proposition 2.39 *A group of order 6 must be isomorphic to either the cyclic group $\mathbb{Z}_6$ or the dihedral group D_3 .*

Proof Let G be a group with six elements. Cauchy's Theorem asserts the existence of an element g of order 2 and an element h of order 3. These are distinct from each other and the identity element e . Each of them generates a cyclic subgroup:

$$\langle g \rangle = \{e, g\} \qquad \langle h \rangle = \{e, h, h^2\}$$

There are therefore six possible elements:

$$G = \{e, g, h, h^2, gh, gh^2\}$$

Now consider the element hg . This element does not belong to the cyclic subgroup $\langle h \rangle$ and it's not equal to g . So we have two possibilities: either $hg = gh$ or $hg = gh^2$.

If $hg = gh$ then this means that every element of the subgroup $\langle g \rangle$ commutes with every element of the subgroup $\langle h \rangle$; furthermore $G = \langle g \rangle \langle h \rangle = \{xy : x \in \langle g \rangle, y \in \langle h \rangle\}$ and $\langle g \rangle \cap \langle h \rangle = \{e\}$. So by Proposition 2.12 we know that $G \cong \langle g \rangle \times \langle h \rangle \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3$. Since 2 and 3 are coprime, Proposition 1.32 tells us that $\mathbb{Z}_2 \oplus \mathbb{Z}_3 \cong \mathbb{Z}_6$, and hence $G \cong \mathbb{Z}_6$.

The other possibility is that $hg = gh^2$; in this case the function $f: G \rightarrow D_3$ given by

$$g \mapsto m_1, \qquad h \mapsto r, \qquad gh \mapsto m_2, \qquad gh^2 \mapsto m_3,$$

is an isomorphism (check this) $G \cong D_3$. $\square$

In Chapter 5 we will prove the more general result that if p is an odd prime then any group of order $2p$ is either dihedral or cyclic; that is, isomorphic to D_p or $\mathbb{Z}_{2p}$.²

² Proposition 5.29, page 153.

We can also, with a bit more effort, classify groups of order 8 up to isomorphism:

Proposition 2.40 *A group of order 8 must be isomorphic to one of the groups $\mathbb{Z}_8$, $\mathbb{Z}_4 \oplus \mathbb{Z}_2$, $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, the dihedral group D_4 or the quaternion group Q_8 .*

The proof follows a similar but slightly more involved pattern to the order-6 case.

Proof Let G be a group with 8 elements. By Proposition 2.34, elements of G must be of order 1 (the identity e), 2, 4 or 8.

If G has an element g of order 8 then $G = \langle g \rangle \cong \mathbb{Z}_8$.

Suppose instead that G has an element g of order 4, but no higher-order elements. Then $\langle g \rangle = \{e, g, g^2, g^3\}$. Let h be some element of $G \setminus \langle g \rangle$ and look at the coset $\langle g \rangle h = \{h, gh, g^2h, g^3h\}$. Hence $G = \langle g \rangle \cup \langle g \rangle h = \{e, g, g^2, g^3, h, gh, g^2h, g^3h\}$.

Now look at the element hg . This clearly isn't a power of g , and is thus not in $\langle g \rangle$. Also, $hg \neq h$, since the right cancellation law (Proposition 1.15) would then imply $g = e$ which we've already decided not to be the case. Furthermore, if $hg = g^2h$ then (by multiplying on the left by h^{-1}) it follows that $g = h^{-1}g^2h$, which implies that

$$g^2 = h^{-1}g^2hh^{-1}g^2h = h^{-1}g^4h = h^{-1}eh = h^{-1}h = e.$$

But g has order 4, so $g^2 \neq e$.

So we're left with two possibilities: either $hg = gh$ or $hg = g^3h$. Also, we have two possibilities for the order of h : either 2 or 4. Notice that $h \in \langle g \rangle h$, since $h \notin \langle g \rangle$. Furthermore, $h^2 \neq g$ since this would require $|h| = 8$; for very similar reasons, $h^2 \neq g^3$. So if h has order 2 then $h^2 = e$, and if h has order 4, the only possibility is that $h^2 = g^2$.

So if G has an element of order 4, we have four possibilities:

- (i) If $hg = gh$ and $|h| = 2$, then G is abelian and the function $f: G \rightarrow \mathbb{Z}_4 \oplus \mathbb{Z}_2$ given by $g \mapsto (1, 0)$ and $h \mapsto (0, 1)$ is an isomorphism $G \cong \mathbb{Z}_4 \oplus \mathbb{Z}_2$.
- (ii) If $hg = g^3h$ and $|h| = 2$, then the function $f: G \rightarrow D_4$ given by $g \mapsto r$ and $h \mapsto m_1$ yields an isomorphism $G \cong D_4$.
- (iii) If $hg = gh$ and $|h| = 4$, then G is abelian. Also, $|gh^{-1}| = 2$ since $gh^{-1}gh^{-1} = g^2h^{-2} = g^2g^{-2} = e$. In this case, the function $f: G \rightarrow \mathbb{Z}_4 \oplus \mathbb{Z}_2$ given by $g \mapsto (1, 0)$ and $gh^{-1} \mapsto (0, 1)$ provides an isomorphism $G \cong \mathbb{Z}_4 \oplus \mathbb{Z}_2$.

- (iv) If $hg = g^3h$ and $|h| = 4$, then the function $f: G \rightarrow Q_8$ given by $g \mapsto I$ and $h \mapsto J$ gives an isomorphism $G \cong Q_8$.

That completes the case where G has an element of order 4. All that remains is to consider the case where all the non-identity elements in G have order 2. In this case, G is abelian, by Proposition 1.20. Choose $g, h, k \in G \setminus \{e\}$ such that $gh \neq k$. The subgroup $\{e, g, h, gh\}$ is isomorphic to the Klein 4-group $V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Let $K = \{e, k\} = \langle k \rangle \cong \mathbb{Z}_2$. Then $HK = G$, the intersection $H \cap K = \{e\}$, and every element of H commutes with both elements of K (since G is abelian). Therefore, by Proposition 2.12 we have

$$G \cong HK \cong H \times K \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2.$$

There are no further cases to consider, so the proof is complete. $\square$

In Section 7.A we will use similar techniques, as well as some more sophisticated ones, to classify all groups of order less than 32.

Mathematics is the queen of sciences and arithmetic the queen of mathematics. She often condescends to render service to astronomy and other natural sciences, but in all relations she is entitled to the first rank.

— Carl Friedrich Gauss (1777–1855),
quoted by Wolfgang Sartorius von
Walterhausen (1809–1876), in *Gauss
zum Gedächtnis* (1856)

2.3 Euler's Theorem and Fermat's Little Theorem

IN THIS SECTION, we examine two applications of Lagrange's Theorem to number theory, and revisit Euler's totient function ϕ .

First we introduce some finite multiplicative groups.

Definition 2.41 Let $\mathbb{Z}_n^\times$ consist of those integers $0, \dots, n-1$ which form a group under modulo- n multiplication. The identity element is clearly 1, and all the other elements are those integers $0 \leq k \leq n-1$ for which there exists a modulo- n multiplicative inverse; that is, an integer $0 \leq h \leq n-1$ where $hk \equiv 1 \pmod{n}$. Clearly 0 is not invertible, so $0 \notin \mathbb{Z}_n^\times$.

In fact, $\mathbb{Z}_n^\times$ consists of exactly those integers $0 \leq k \leq n-1$ for which $\gcd(k, n) = 1$.

This is a special case of a construct we'll meet later on in the second part of the book: the **group of units** of a ring.³

Example 2.42 In the following list, $\mathbb{Z}_n^\times$ is the multiplicative group described above, while $\mathbb{Z}_n$ denotes the additive group of integers modulo n .

$$\begin{array}{ll} \mathbb{Z}_2^\times = \{1\} & \mathbb{Z}_3^\times = \{1, 2\} \cong \mathbb{Z}_2 \\ \mathbb{Z}_4^\times = \{1, 3\} \cong \mathbb{Z}_2 & \mathbb{Z}_5^\times = \{1, 2, 3, 4\} \cong \mathbb{Z}_4 \\ \mathbb{Z}_6^\times = \{1, 5\} \cong \mathbb{Z}_2 & \mathbb{Z}_7^\times = \{1, 2, 3, 4, 5, 6\} \cong \mathbb{Z}_6 \\ \mathbb{Z}_8^\times = \{1, 3, 5, 7\} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 & \mathbb{Z}_9^\times = \{1, 2, 4, 5, 7, 8\} \cong \mathbb{Z}_6 \end{array}$$

³ Proposition 8.34, page 332.

We can see that $\mathbb{Z}_5^\times$ is isomorphic to $\mathbb{Z}_4$ and not $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ by observing that the element $2 \in \mathbb{Z}_5^\times$ has multiplicative order 4, since $2^2 = 4 \not\equiv 1 \pmod{5}$ but $2^4 = 16 \equiv 1 \pmod{5}$. On the other hand, in $\mathbb{Z}_8^\times$ we find that $3^2 = 9 \equiv 1 \pmod{8}$, $5^2 = 25 \equiv 1 \pmod{8}$ and $7^2 = 49 \equiv 1 \pmod{8}$, so there are no elements of order 4, and hence $\mathbb{Z}_8^\times$ is isomorphic to the Klein group $V_4 = \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Similar arguments can be applied to prove the other isomorphisms listed.

The following lemma tells us useful information about the group $\mathbb{Z}_n^\times$:

Lemma 2.43 *Let k be a non-negative integer such that $0 < k < n$. Then $k \in \mathbb{Z}_n^\times$ if and only if $\gcd(k, n) = 1$; that is, if k and n are coprime.*

Proof If k and n are not coprime, then $\gcd(k, n) = d > 1$. So d divides km for any $m \in \{0, \dots, n-1\}$ and hence $km \neq 1 \in \mathbb{Z}_n^\times$. Thus k has no inverse modulo n , and is hence not an element of $\mathbb{Z}_n^\times$.

If, on the other hand, k and n are coprime, then by the argument in Proposition 1.26 and the following paragraph, k generates $\mathbb{Z}_n$. Hence there exists $m \in \mathbb{Z}$ such that $mn = 1 \in \mathbb{Z}_n$. Or, equivalently, $mk = an + 1 \in \mathbb{Z}$. Now let $l = [m]_n$, the **remainder** or **residue** of m modulo n . Then $lk = bn + 1 \in \mathbb{Z}$ for some $b \in \mathbb{Z}$, hence $lk = 1 \in \mathbb{Z}_n^\times$ and so k is an invertible element with inverse l , and is thus in $\mathbb{Z}_n^\times$. $\square$

This tells us that $\mathbb{Z}_n^\times$ consists of exactly those non-negative integers $k \in \mathbb{Z}_n = \{0, \dots, n-1\}$ where $\gcd(k, n) = 1$. Thus $|\mathbb{Z}_n^\times|$ is the number of these coprime integers, for which concept we already have a name:

Definition 2.44 **Euler's totient function** $\phi(n)$ is the number of integers $k \in \{0, \dots, n-1\}$ such that $\gcd(k, n) = 1$:

$$\phi(n) = |\{k \in \mathbb{Z} : 0 \leq k < n \text{ and } \gcd(k, n) = 1\}|$$

We will study $\phi(n)$ further later on, but for the moment we now have enough to state and prove Euler's Theorem:

Theorem 2.45 (Euler's Theorem) *Suppose k and n are two coprime, positive integers. Then*

$$k^{\phi(n)} \equiv 1 \pmod{n}$$

or, equivalently, $n \mid (k^{\phi(n)} - 1)$.

Proof Lemma 2.43 tells us that $|\mathbb{Z}_n^\times| = \phi(n)$ since $k \in \mathbb{Z}_n^\times$ if and only if $\gcd(k, n) = 1$. Let $[k]_n$ denote the remainder of k modulo n . Then clearly $[k]_n \in \mathbb{Z}_n^\times$ and so, by Proposition 2.34, the order $|[k]_n|$ of $[k]_n$ in $\mathbb{Z}_n^\times$ must be a factor of $|\mathbb{Z}_n^\times| = \phi(n)$. So there exists some positive integer $l \in \mathbb{Z}$ such that $l|[k]_n| = \phi(n)$. Hence

$$[k]_n^{\phi(n)} = [k]_n^{l|[k]_n|} = ([k]_n^{|[k]_n|})^l = 1^l = 1.$$

So $k^{\phi(n)} \equiv 1 \pmod{n}$. $\square$



Wikimedia Commons / François de Poilly (1623–1693)

The French lawyer and amateur mathematician **Pierre de Fermat** (1601–1665) is chiefly remembered for his work in analytic geometry and number theory, in particular his famous “Last Theorem”. This conjecture, which states that there exist no integer solutions to the equation

$$x^n + y^n = z^n$$

for $n > 2$, remained unproven until 1994, when Andrew Wiles proved the Taniyama–Shimura Conjecture on rational elliptic curves and modular forms, which was known to imply Fermat’s conjecture as a corollary.

The original statement of the conjecture occurs in Fermat’s hand-annotated copy of the treatise *Arithmetica*, written in the third century AD by the Greek mathematician Diophantus of Alexandria. Fermat notes that he has discovered a marvellous proof (“*demonstrationem mirabilem*”) of the insolubility of the general case, but that the margin is too small to contain it (“*hanc marginis exiguitas non caperet*”).

Born in Gascony, the son of a merchant named Dominique Fermat, he studied law at the University of Orléans, and then embarked on a legal career. In 1631, he was appointed as a councillor and judge at the Parlement de Toulouse, thereby entitling him to add the honorific ‘de’ to his surname.

Fluent in six languages (French, Latin, Greek, Occitan, Spanish and Italian), he was also an accomplished writer and poet. He communicated most of his mathematical discoveries informally, in letters to friends, usually omitting the proofs, which were reconstructed by other mathematicians after his death.

In general, $\phi(n)$ is somewhat complicated, and its calculation involves the prime factorisation of n . In the case of a prime integer p , however, $\phi(p)$ has a much simpler form:

Proposition 2.46 *If $p \in \mathbb{N}$ is prime, then $\phi(p) = p - 1$.*

Proof Recall that $\phi(p)$ is the number of integers $k \in \{1, \dots, p-1\}$ for which $\gcd(p, k) = 1$. But since p is prime, all such integers k have this property, and so $\phi(p) = |\{1, \dots, p-1\}| = p - 1$. $\square$

The next proposition is a clever application of Lagrange’s Theorem.

Proposition 2.47 *Suppose $n \geq 3$. Then $\phi(n)$ is even.*

Proof Let $G = \mathbb{Z}_n^\times$ and let $H = \{1, n-1\}$. It is simple to show that H is a subgroup of G : it is closed under multiplication modulo n , with the only nonimmediate case being

$$(n-1)(n-1) = n^2 - 2n + 1 \equiv 1 \pmod{n}.$$

By Lagrange’s Theorem, $|H| = 2$ must be a factor of $|G| = |\mathbb{Z}_n^\times| = \phi(n)$, and so $\phi(n)$ is even. (If $n = 2$ this fails, because $n-1 = 1$ and hence the group $H = \{1\}$ has only one element; indeed, we saw in Example 2.42 that $|\mathbb{Z}_2^\times| = \phi(2) = 1$.) $\square$

Corollary 2.48 (Fermat’s Little Theorem) *Let $p \in \mathbb{N}$ be prime. Then for any $k \in \mathbb{Z}$,*

$$k^p \equiv k \pmod{p}$$

or, equivalently, $p \mid (k^p - k)$.

Proof Proposition 2.46 tells us that $\phi(p) = p-1$, and then by Euler’s Theorem we have

$$k^p = k(k^{p-1}) = k(k^{\phi(p)}) \equiv k \pmod{p}$$

as required. $\square$

These two results illustrate the important connections between algebra and number theory, and demonstrate how versatile algebraic techniques can be when applied to other topics. We will meet a number of other applications over the rest of the book.

Summary

In this chapter, we studied **subgroups**, smaller groups neatly embedded inside larger ones. More precisely, a **subgroup** $H < G$ is a subset $H \subseteq G$ which forms a group in its own right under the same operation (or, strictly speaking, the **induced operation** $*_H$ formed by restricting $*_G$ to $H \times H$).⁴ In order for this to happen, we require that H must be closed under $*_H$ and that H contain inverses of all its elements. A subgroup will often inherit certain properties from the larger group: in particular, subgroups of abelian groups are also abelian,⁵ and subgroups of cyclic groups are also cyclic.⁶

⁴ Definition 2.1, page 40.

⁵ Proposition 2.2, page 41.

⁶ Proposition 2.7, page 43.

We saw a number of examples of subgroups. Every nontrivial group has at least two: itself and the **trivial subgroup** $\{e\}$ consisting of just the identity element. The dihedral group D_n contains the rotation subgroup R_n , which is isomorphic to the cyclic group $\mathbb{Z}_n$. The multiplicative group $\mathbb{R}_*^\times$ of nonzero real numbers⁷ and the **circle group** $C = \{e^{i\theta} : 0 \leq \theta < 2\pi\} \cong U_1$ ⁸ both fit neatly inside the multiplicative group $\mathbb{C}_*^\times$ of nonzero complex numbers. Most of the matrix groups we met in Section 1.2 fit together in interesting ways, for example, the special orthogonal group $SO_n(\mathbb{R})$ is a subgroup of the orthogonal group $O_n(\mathbb{R})$, which in turn is a subgroup of the general linear group $GL_n(\mathbb{R})$, as are the special linear group $SL_n(\mathbb{R})$ and the symplectic group $Sp_{2n}(\mathbb{R})$.⁹ For any positive integer n , the group $\langle n \rangle = n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$ itself; more generally for any element g of a group G , the set $\langle g \rangle = \{g^i : i \in \mathbb{Z}\}$ is a subgroup of G , called the **cyclic subgroup** generated by g .¹⁰ We can often describe the relationship between a group and its various subgroups, and their relationships to each other, by means of a **lattice diagram**; in particular it turns out that the intersection of any two subgroups will also be a subgroup.¹¹

⁷ Example 2.4, page 41.

⁸ Example 2.5, page 42.

⁹ Example 2.6, page 42.

¹⁰ Proposition 2.8, page 44.

¹¹ Proposition 2.9, page 45.

Permutation groups, that is, subgroups of (finite or infinite) symmetric groups are particularly important: Cayley's Theorem tells us that any group is isomorphic to a group of this type.¹² This is true for finite groups like $\mathbb{Z}_3 \cong \{\iota, (1\ 2\ 3), (1\ 3\ 2)\}$ ¹³ and also for infinite groups like $\mathbb{Z}$.¹⁴

¹² Theorem 2.13, page 47.

¹³ Example 2.14, page 48.

¹⁴ Example 2.15, page 49.

In the proof of Cayley's Theorem we introduced a class of permutations $\lambda_g: G \rightarrow G$ for all $g \in G$, and shortly afterwards we examined the effect of these permutations on some subgroup H of G . This led quite naturally to the concept of a **left coset**¹⁵

¹⁵ Definition 2.16, page 50.

$$gH = \lambda_g(H) = \{g * h : h \in H\}.$$

Analogously, we can define the **right coset**¹⁶

¹⁶ Definition 2.17, page 50.

$$Hg = \rho_g(H) = \{h * g : h \in H\}$$

which, if G is nonabelian, need not be the same as the corresponding left coset.

We saw that two right cosets Hg and Hk must be either equal or completely disjoint,¹⁷ and that they must have the same cardinality (both of these statements are true for left cosets as well).¹⁸ This means that we can use a subgroup H to neatly partition a group G into a collection of equal-sized subsets;¹⁹ from this observation we get Lagrange's Theorem, which says that if H is a subgroup of a finite group G , then the number of elements in H must be a factor of the number of elements in G .²⁰ The number of (left or right) cosets of H in G gives an indication of the relative size of G and H , and is given a special name: we call it the **index** of H in G and denote it $|G : H|$.²¹ The index is multiplicative, in the sense that if $K < H < G$, then $|G : K| = |G : H||H : K|$.²²

Lagrange's Theorem yields an important necessary condition for a subset to be a subgroup; for example we can see straight away that $\mathbb{Z}_{10}$ can have no three-element subgroups since $3 \nmid 10$. It also tells us, almost immediately, that a group of prime order (such as the cyclic group $\mathbb{Z}_p$) can have no subgroups apart from itself and the trivial subgroup $\{e\}$.²³ With a little more thought, we can show that the order $|g|$ of an *element* $g \in G$ must divide the order $|G|$ of the group G , since g generates a cyclic subgroup $\langle g \rangle$ which must have a number of elements equal to $|g|$.²⁴ From here it's only a short step to prove that if p is prime, then there can only be one group (up to isomorphism) with p elements, namely $\mathbb{Z}_p$ itself.²⁵

The condition we get from Lagrange's Theorem is necessary but not sufficient: it says nothing to guarantee the existence of subgroups of a particular order, and indeed such a subgroup might not exist. For example, we saw that although 6 is a factor of $12 = |A_4|$, it so happens that A_4 has no subgroup of order 6.²⁶

We do, however, have a number of theorems which provide partial converses to Lagrange's Theorem. The three Sylow Theorems are a bit more advanced, and we'll discuss them at length in Chapter 7, but Cauchy's Theorem is more straightforward. This states that if p is a *prime* factor of the order $|G|$ of a finite group G , then G will contain a subgroup of order p .²⁷

We used Lagrange's Theorem and Cauchy's Theorem to classify (that is, write down a complete list of) groups with up to eight elements, up to isomorphism.

There is one group of order 1: the trivial group $\{e\}$. There is one group each of orders 2, 3, 5 and 7, namely the cyclic groups $\mathbb{Z}_2$, $\mathbb{Z}_3$, $\mathbb{Z}_5$ and $\mathbb{Z}_7$.²⁸ There are two groups of order 4: the cyclic group $\mathbb{Z}_4$ and the

¹⁷ Corollary 2.27, page 54.

¹⁸ Proposition 2.29, page 54.

¹⁹ Corollary 2.28, page 54.

²⁰ Theorem 2.30, page 54.

²¹ Definition 2.31, page 55.

²² Proposition 2.32, page 55.

²³ Corollary 2.33, page 55.

²⁴ Proposition 2.34, page 55.

²⁵ Proposition 2.35, page 56.

²⁶ Example 2.36, page 56.

²⁷ Theorem 2.37, page 57.

²⁸ Proposition 2.35, page 56.

Klein 4-group V_4 .²⁹ There are two groups of order 6: the cyclic group $\mathbb{Z}_6$ and the dihedral group D_3 (which happens to be isomorphic to the symmetric group S_3).³⁰ Finally, there are five groups of order 8: the cyclic group $\mathbb{Z}_8$, together with the direct sums $\mathbb{Z}_4 \oplus \mathbb{Z}_2$ and $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, the quaternion group Q_8 and the dihedral group D_4 .³¹

In Section 2.3 we looked at a few applications of Lagrange's Theorem to number theory. First we investigated some finite multiplicative groups of integers: $\mathbb{Z}_n^\times$ consists of those integers $1, \dots, n-1$ which form a group under modulo- n multiplication.^{32,33} With a little thought, we see that these are exactly those integers $0 < k < n$ that are coprime to n , or equivalently for which $\gcd(k, n) = 1$.³⁴ The number of such integers for a given n , that is, the number of elements in $\mathbb{Z}_n^\times$ is given by Euler's **totient function** $\phi(n)$.³⁵ Euler's Theorem says that if k and n are coprime then $k^{\phi(n)} \equiv 1 \pmod{n}$, or equivalently that n divides $(k^{\phi(n)} - 1)$.³⁶

The totient function $\phi(n)$ is quite complicated in general, and we'll study it in a little more detail later, but we can already state a couple of useful facts about it. Firstly, if p is prime, then $\phi(p) = p - 1$.³⁷ Using this together with Euler's Theorem we obtain Fermat's Little Theorem, which states that for any integer k we have $k^p \equiv k \pmod{p}$, or equivalently that p divides $(k^p - k)$.³⁸ Secondly, if $n \geq 3$ then $\{1, n-1\}$ forms a 2-element subgroup of $\mathbb{Z}_n^\times$. Lagrange's Theorem then tells us that $|\mathbb{Z}_n^\times|$ must have 2 as a factor, which is equivalent to saying that $\phi(n)$ is even for $n \geq 3$.³⁹

²⁹ Proposition 2.38, page 59.

³⁰ Proposition 2.39, page 60.

³¹ Proposition 2.40, page 61.

³² Definition 2.41, page 62.

³³ Example 2.42, page 62.

³⁴ Lemma 2.43, page 63.

³⁵ Definition 2.44, page 63.

³⁶ Theorem 2.45, page 63.

³⁷ Proposition 2.46, page 64.

³⁸ Theorem 2.48, page 64.

³⁹ Proposition 2.47, page 64.

References and further reading

Fermat's Last Theorem is one of the most celebrated problems in mathematics, its deceptively simple statement nevertheless resisting all attempts at a solution for about three and a half centuries. There are a number of readable accounts of its history and ultimate solution; probably the best are the following books by Amir Aczel and Simon Singh.

- A D Aczel, *Fermat's Last Theorem*, Penguin (1997)
- S Singh, *Fermat's Last Theorem*, Fourth Estate (1997)

If you can find a copy, Singh's 1996 BBC *Horizon* documentary *Fermat's Last Theorem* is also worth watching: at the time of writing, it is available to watch online via the BBC's website.

An interesting account of the history and subsequent development of Cauchy's Theorem can be found in the following article:

- M Meo, *The mathematical life of Cauchy's group theorem*, *Historia Mathematica* 31 (2004) 196–221

Cauchy's original exposition and proof ran to about nine pages at the end of a long article on permutation groups published in 1845.

- A L Cauchy, *Mémoire sur les arrangements que l'on peut former avec des lettres données, et sur les permutations et substitutions à l'aide desquelles on passe d'un arrangement à un autre*, Exercices d'analyse et de physique mathématique 3 (1845) 151–252

Meo also notes that Cauchy's proof contains a subtle error: at one point he assumes that the direct product of two subgroups is also a subgroup. This is true if both subgroups are **normal**, as we will see in Corollary 3.23 in the next chapter, but not in general.

An elegant and concise proof, just ten lines long, was published in 1959 by the American mathematician James McKay (1928–2012):

- J H McKay, *Another proof of Cauchy's group theorem*, American Mathematical Monthly 66.2 (1959) 119

Exercises

2.1 For each of the following groups G and subsets $H \subseteq G$, say whether or not H is a subgroup of G ; if not, explain briefly why not.

(a) $G = \mathbb{Z}, H = \mathbb{N}$.	(g) $G = D_5, H = \{e, r^2\}$.
(b) $G = \mathbb{Z}, H = \{0\}$.	(h) $G = D_5, H = \{e, m_3\}$.
(c) $G = \mathbb{Z}, H = \{-1, 0, 1\}$.	(i) $G = GL_2(\mathbb{R}), H = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} : \right.$
(d) $G = \mathbb{Z}, H = 7\mathbb{Z}$.	$\left. a, b, c \in \mathbb{R}, ac \neq 0 \right\}$.
(e) $G = S_3, H = \{(123), (132)\}$.	(j) $G = GL_2(\mathbb{R}), H = GL_2(\mathbb{Q})$.
(f) $G = S_3, H = \{i, (12)\}$.	(k) $G = \mathbb{Z} \oplus \mathbb{Z}, H = \{(a, b) : a+b \text{ is even}\}$.

2.2 Repeat Example 2.14 with the cyclic group $\mathbb{Z}_4$ and the Klein group $V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

2.3 Let $m, n \in \mathbb{N}$ such that $n \geq 2$ and $0 \leq m \leq n$. Let H be the subgroup of the symmetric group S_n consisting of those permutations which permute the first m elements among themselves. Use Lagrange's Theorem in a manner analogous to Proposition 2.47 to show that the binomial coefficient $\binom{n}{m} = \frac{n!}{m!(n-m)!}$ is an integer.

2.4 Find the orders of $\mathbb{Z}_n^\times$ for $n = 10, 12, 21, 60$. In the cases where $|\mathbb{Z}_n^\times| = 4$, determine whether $\mathbb{Z}_n^\times$ is isomorphic to $\mathbb{Z}_4$ or the Klein group V_4 .

2.5 Find the subgroups of $\mathbb{Z}_{14}^\times$.

2.6 Let p be prime, and show that $(p-1)$ is the only element of $\mathbb{Z}_p^\times$ of order 2. By considering the product of all elements of $\mathbb{Z}_p^\times$, prove Wilson's Theorem: that $(p-1)! \equiv -1 \pmod{p}$ if and only if p is prime.

2.7 Let H be a subgroup of a group G , and suppose that $g \in G$ is some arbitrary element. Show that there exists some element $k \in G$ such that $gH = Hk$.

2.8 (a) Show that a group of even order has at least one element of order 2.

(b) Show that a group of odd order has no elements of order 2.

2.9 (a) If G is abelian, show that $H = \{g \in G : g^2 = e\}$ is a subgroup of G .

(b) Find an example of a nonabelian group G such that H is not a subgroup of G .

2.10 (a) If G is abelian, show that $H = \{g^2 : g \in G\}$ is a subgroup of G .

(b) Find an example of a nonabelian group G for which H is a subgroup.

(c) Find a nonabelian group G for which H is not a subgroup.

3 Normal subgroups

IN THE LAST CHAPTER we introduced the concept of a **coset** of a subgroup H of a group G . In the case where G is abelian, obviously each left coset gH is equal to the corresponding right coset Hg , but if G is nonabelian this need not be the case, and often isn't. In this chapter we will examine this topic further, along the way deriving a condition on H that forces the left cosets to coincide with their right coset twins. Subgroups H which have this property are called **normal subgroups** and have an important rôle to play: given a normal subgroup H of a group G we can construct the **quotient group** G/H . This method of decomposing larger groups into products of smaller ones gives us new techniques for understanding their internal structure. This is (very loosely) analogous to the number-theoretic study of the multiplicative properties of integers by factorising them into smaller ones, and just as we then find a class of integers, namely the prime numbers, which can't be factorised into smaller ones, we will also investigate the class of **simple groups**: those which have no normal subgroups (apart from themselves and the trivial subgroup) and hence can't be factorised into smaller ones.

3.1 Cosets and conjugacy classes

IT IS AN INTERESTING QUESTION to ask when the left and right cosets of a subgroup $H \leq G$ coincide: is there some property that H or G might have that then forces $gH = Hg$ for any element $g \in G$?

An obvious sufficient condition is that G (and hence, by Proposition 2.2, H) be abelian, but we know already that this isn't a necessary condition: at the beginning of Section 2.2 we saw that the rotation subgroup $R_3 \leq D_3$ has this property, but D_3 isn't abelian, although $R_3 \cong \mathbb{Z}_3$ is. Perhaps the subgroup H has to be abelian in order to have this coincidence property. This is a reasonable suggestion, but the next example

Up and after ordering some things towards my wife's going into the country, to the office, where I spent the morning upon my measuring rules very pleasantly till noon, and then comes Creed and he and I talked about mathematiques, and he tells me of a way found out by Mr. Jonas Moore which he calls duodecimal arithmetique, which is properly applied to measuring, where all is ordered by inches, which are 12 in a foot, which I have a mind to learn.

— Samuel Pepys (1633–1703),
diary entry for Tuesday 9 June 1663

The *species* and *genus* are always the work of nature; the *variety* often that of culture; and the *class* and *order* are the work of nature and art.

— Carl Linnaeus (1707–1778),
Philosophia Botanica (1751) 162

shows that it's wrong.

Example 3.1 The symmetric group S_4 contains the alternating group A_4 as a subgroup. The left cosets of A_4 in S_4 are

$$\begin{aligned} A_4 &= \iota A_4 = (123)A_4 = (132)A_4 = (124)A_4 = (142)A_4 \\ &= (134)A_4 = (143)A_4 = (234)A_4 = (243)A_4 \\ &= (12)(34)A_4 = (13)(24)A_4 = (14)(23)A_4 \\ &= \{\iota, (123), (132), (124), (142), (134), (143), (234), (243), \\ &\quad (12)(34), (13)(24), (14)(23)\} \end{aligned}$$

$$\begin{aligned} S_4 \setminus A_4 &= (12)A_4 = (13)A_4 = (14)A_4 = (23)A_4 = (24)A_4 = (34)A_4 \\ &= (1234)A_4 = (1243)A_4 = (1324)A_4 \\ &= (1342)A_4 = (1423)A_4 = (1432)A_4 \\ &= \{(12), (13), (14), (23), (24), (34), \\ &\quad (1234), (1243), (1324), (1342), (1423), (1432)\} \end{aligned}$$

while the right cosets of A_4 in S_4 are

$$\begin{aligned} A_4 &= A_4 \iota = A_4(123) = A_4(132) = A_4(124) = A_4(142) \\ &= A_4(134) = A_4(143) = A_4(234) = A_4(243) \\ &= A_4(12)(34) = A_4(13)(24) = A_4(14)(23) \\ &= \{\iota, (123), (132), (124), (142), (134), (143), (234), (243), \\ &\quad (12)(34), (13)(24), (14)(23)\} \end{aligned}$$

$$\begin{aligned} S_4 \setminus A_4 &= A_4(12) = A_4(13) = A_4(14) = A_4(23) = A_4(24) = A_4(34) \\ &= A_4(1234) = A_4(1243) = A_4(1324) \\ &= A_4(1342) = A_4(1423) = A_4(1432) \\ &= \{(12), (13), (14), (23), (24), (34), \\ &\quad (1234), (1243), (1324), (1342), (1423), (1432)\} \end{aligned}$$

These are clearly the same, although neither A_4 nor S_4 are abelian.

It turns out that the answer is a bit more complicated, and in order to understand it, we need to introduce a new concept: that of **conjugacy**. Impatient or alert readers may wonder why it matters whether the left and right cosets are the same. This question will be answered in the next section, where we'll find that subgroups having this property are useful for understanding the internal structure of the larger group.

Definition 3.2 An element $g \in G$ is **conjugate** to an element $h \in G$ if there exists some element $k \in G$ such that $k * g * k^{-1} = h$. We say also that h is the result of **conjugating** g with k .

This may seem like a strange criterion to impose on elements of a group, but if you've studied a bit of linear algebra you should have

met this concept before, although probably under a different name.

Definition 3.3 Two $n \times n$ square matrices A and B are said to be **similar** if there exists an invertible $n \times n$ matrix P such that $B = PAP^{-1}$.

This is exactly the same idea as conjugacy within a matrix group such as $GL_n(\mathbb{R})$.

We'll stay with similar matrices for a little while, and think carefully about what this means. An $n \times n$ matrix A with elements in $\mathbb{R}$ can be regarded as representing a linear map $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$. But there are many (uncountably infinitely many, in fact) other matrices which also represent that same linear map, so what's going on in this context is that A represents f relative to some choice of **basis** or **coordinate system** for $\mathbb{R}^n$. The equation $B = PAP^{-1}$ says that both A and B represent the same linear map $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ relative to different bases for $\mathbb{R}^n$, and P is the **change of basis** matrix that translates from one basis (or coordinate system) to the other.

More precisely, let $S = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ and $T = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ be two bases for $\mathbb{R}^n$, let A be the matrix representing f relative to the basis S , let B be the matrix representing f relative to the basis T , and let P be the change of basis matrix from S to T .

The point of P is that given any vector $\mathbf{w} \in \mathbb{R}^n$, expressed as an n -element column vector relative to the basis S , the vector $P\mathbf{w}$ will be an n -element column vector representing the same vector in terms of the basis T .

The equation $B = PAP^{-1}$, then, tells us how A and B are connected: given a matrix A representing f relative to the basis S , in order to work out the matrix representing f relative to the basis T , we have to first map from T coordinates to S coordinates (using the inverse P^{-1}), apply A (in S coordinates) and then map back into T coordinates again (using P). This procedure gives PAP^{-1} , which must therefore be equal to B .

We can regard this matrix P as representing a change of coordinate system, but any such matrix has to be invertible, and hence an element of $GL_n(\mathbb{R})$. So another way of looking at all this is to say that we can rewrite the element $B \in GL_n(\mathbb{R})$ as a product of three other elements of $GL_n(\mathbb{R})$. In the case of $GL_n(\mathbb{R})$, the element P has a nice interpretation as a change of coordinate system, but this isn't always the case (even in analogy) when we look at other groups.

For example, let's consider the dihedral group D_3 again. The element m_1 has a geometric interpretation as a reflection in one of the axes of symmetry of an equilateral triangle. But so does m_2 , just through

	e	r	r^2	m_1	m_2	m_3
e	e	e	e	e	e	e
r	r	r	r	r^2	r^2	r^2
r^2	r^2	r^2	r^2	r	r	r
m_1	m_1	m_2	m_3	m_1	m_3	m_2
m_2	m_2	m_3	m_1	m_3	m_2	m_1
m_3	m_3	m_1	m_2	m_2	m_1	m_3

Table 3.1: Conjugation in D_3

The subtables

	e
e	e
	r r^2
r	r r
r^2	r^2 r^2
	m_1 m_2 m_3
m_1	m_1 m_3 m_2
m_2	m_3 m_2 m_1
m_3	m_2 m_1 m_3

are examples of another interesting algebraic structure: a **quandle**. These objects (and a slightly more general structure called a **rack**) occur naturally in the study of knot theory; the third, for example, represents the nontrivial 3-colouring of the trefoil knot.

one of the other axes. Thinking about it, we can get the same result as applying m_2 by rotating the whole triangle by a third of a rotation, doing m_1 and then rotating the triangle back again. So m_1 and m_2 are conjugate to each other. On the other hand, m_1 and r aren't conjugate to each other, since there exists no $g \in D_3$ such that $m_1 = grg^{-1}$.

Table 3.1 shows this more clearly. The group element at the intersection of row g with column h is that obtained by conjugating g with h , that is, hgh^{-1} . Here we see that e is conjugate only to itself, r is conjugate to r^2 and vice versa, and the reflections m_1 , m_2 and m_3 are all conjugate to each other.

The operation of conjugation (or, in fact, the relation “is conjugate to”) neatly partitions D_3 into three subsets (which, with the exception of $\{e\}$, are not subgroups). The obvious next question is: does this happen in general? Does the relation “is conjugate to” always partition a group into non-overlapping subsets? The answer is yes, because of the following fact:

Proposition 3.4 *Conjugacy is an equivalence relation.*

Proof To show that conjugacy is an equivalence relation we need to show that it is reflexive, symmetric and transitive.

Conjugacy is reflexive, since for any element $g \in G$ we have $e * g * e^{-1} = g$ and hence $g \sim g$.

Conjugacy is symmetric, since if $g \sim h$ there exists some $x \in G$ such that $h = x * g * x^{-1}$. Multiplying both sides of this equation on the left by x^{-1} and on the right by x yields

$$x^{-1} * h * x = x * x^{-1} * g * x^{-1} * x = e * g * e = g.$$

So, since $x^{-1} \in G$ we have $g = (x^{-1}) * h * (x^{-1})^{-1}$ and hence $h \sim g$.

Finally, suppose that $g \sim h$ and $h \sim k$. So there exists $x \in G$ such that $h = x * g * x^{-1}$, and there exists $y \in G$ such that $k = y * h * y^{-1}$. Substituting the first of these into the second we get

$$k = y * h * y^{-1} = y * x * g * x^{-1} * y^{-1} = (y * x) * g * (y * x)^{-1}.$$

So g is conjugate to k via $(y * x) \in G$ and hence $g \sim k$. Therefore conjugacy is transitive. $\square$

We call the equivalence classes of this relation **conjugacy classes**.

Example 3.5 The conjugacy classes of D_3 are

$$\{e\}, \quad \{r, r^2\} \quad \text{and} \quad \{m_1, m_2, m_3\}.$$

Conjugacy in a matrix group is the same as similarity between matrices. The next example characterises the conjugacy classes in $GL_n(\mathbb{R})$.

Example 3.6 We know from linear algebra that some, but not all, matrices in $GL_n(\mathbb{R})$ are similar to (or, in our more general terminology, conjugate to) a diagonal matrix. More generally, every matrix $M \in GL_n(\mathbb{R})$ is conjugate to one in **Jordan canonical form**. So, the conjugacy classes of $GL_n(\mathbb{R})$ are indexed by the various types of $n \times n$ Jordan matrices.

A **Jordan block** with eigenvalue $\lambda \in \mathbb{R}$ of degree k is a $k \times k$ matrix $J_{\lambda,k} = [a_{ij}]$ with diagonal elements $a_{ii} = \lambda$, the supradiagonal elements $a_{i,i+1} = 1$ for $1 \leq i < k$ and all other elements $a_{ij} = 0$ if $j \neq i$ or $i+1$.

A matrix is in **Jordan canonical form** if it can be decomposed as a block diagonal matrix where the diagonal blocks are Jordan blocks, and all the off-diagonal blocks are zero.

The conjugacy classes of matrices in $GL_n(\mathbb{R})$ are in bijective correspondence with the different Jordan matrices.

So, considering $GL_2(\mathbb{R})$ we have three different types:

$$\text{diag}(J_{\lambda,1}, J_{\lambda,1}) = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad \text{diag}(J_{\lambda,1}, J_{\mu,1}) = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \quad J_{\lambda,2} = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$$

where $\lambda, \mu \in \mathbb{R}$ and $\lambda \neq \mu$. Thus $GL_2(\mathbb{R})$ has uncountably many conjugacy classes, but each is one of these three types. The group $GL_3(\mathbb{R})$ has uncountably many conjugacy classes too, which may be divided into six different types. As before, the 3×3 identity matrix I_3 is in a conjugacy class on its own.

Example 3.7 The conjugacy classes of $\mathbb{Z}$ are singleton sets, each consisting of a single integer:

$$\dots, \{-2\}, \{-1\}, \{0\}, \{1\}, \{2\}, \dots$$

To see this, consider an arbitrary integer $n \in \mathbb{Z}$. Conjugating this with any other integer $m \in \mathbb{Z}$ yields $m + n + (-m) = n$. So n can only be conjugate to itself.

This is true for any abelian group, as the following proposition shows.

Proposition 3.8 *Let G be an abelian group. Then for any element $g \in G$, the conjugacy class of g is $\{g\}$.*

Proof Conjugating g by another element $h \in G$ gives

$$h * g * h^{-1} = h * h^{-1} * g = e * g = g.$$

So g is only conjugate to itself, and thus its conjugacy class is $\{g\}$. $\square$

While we're on the subject of conjugation, it's worth noting that the conjugation operation can be used to construct an isomorphism from a group to itself:

Proposition 3.9 *Let $g \in G$ be some fixed element of a group G . Then the function $f_g: G \rightarrow G$, defined by $h \mapsto g * h * g^{-1}$, is an isomorphism from G to itself.*

Proof This function f_g is bijective: an inverse is given by the function $f_g^{-1} = f_{g^{-1}}: G \rightarrow G$ which maps h to $g^{-1} * h * g$.

Alternatively, to show that f_g is injective, suppose that $f_g(h) = f_g(k)$ for some $h, k \in G$. Then $g * h * g^{-1} = g * k * g^{-1}$, and multiplying on the left by g^{-1} and on the right by g we obtain $h = k$ as required. To show surjectivity, consider an element $h \in G$. Then

$$f_g(g^{-1} * h * g) = g * g^{-1} * h * g * g^{-1} = e * h * e = h.$$

It satisfies the structure condition too, since for any $h, k \in G$ we have

$$\begin{aligned} f_g(h * k) &= g * h * k * g^{-1} = g * h * e * k * g^{-1} = \\ &= g * h * g^{-1} * g * k * g^{-1} = f_g(h) * f_g(k) \end{aligned}$$

as required. □

An isomorphism from a group to itself is called an **automorphism**, and automorphisms of the form discussed in Proposition 3.9 are called **inner automorphisms**.

We are now in a position to study precisely what conditions a subgroup $H \leq G$ must satisfy in order to ensure that each of its left cosets is equal to the corresponding right coset.

Looking again at the dihedral group D_3 we recall that the rotation subgroup $R_3 = \{e, r, r^2\}$ has this property $gR_3 = R_3g$ for any $g \in D_3$. The subgroups $\{e, m_1\}$, $\{e, m_2\}$ and $\{e, m_3\}$ don't, however, although the trivial subgroup $\{e\}$ and the full group D_3 do.

The clue lies in the observation that $D_3 = \{e\} \cup \{r, r^2\} \cup \{m_1, m_2, m_3\}$, $R_3 = \{e\} \cup \{r, r^2\}$ and the trivial subgroup $\{e\}$ are all unions of conjugacy classes of elements of D_3 , while the reflection subgroups $\{e, m_1\}$, $\{e, m_2\}$ and $\{e, m_3\}$ aren't.

Why does this matter? Well, if a subgroup $H < G$ is a union of conjugacy classes (which not every subgroup need be) then it ensures that for any element $h \in H$, all elements of the form $g * h * g^{-1}$ are also in H . That is, H is closed under conjugation by elements of G .

The following proposition confirms that we're on the right track:

Proposition 3.10 *A subgroup $H < G$ is a union of conjugacy classes if and only if $Hg = gH$ for all $g \in G$.*

Proof Suppose H is a union of conjugacy classes. Then for any $g \in G$ and $h \in H$ we know that $g * h * g^{-1}$ and $g^{-1} * h * g$ must belong to H . The element $g * h$ clearly belongs to the left coset gH , but since

$g * h = (g * h * g^{-1}) * g$ it must also lie in the right coset Hg , and so $gH \subseteq Hg$. Similarly, $h * g$ lies in the right coset Hg , but since $h * g = g * (g^{-1} * h * g)$, it must also lie in the left coset gH , and so $Hg \subseteq gH$. Therefore $Hg = gH$.

Now suppose that $gH = Hg$ for all $g \in G$. If $h \in H$ then its conjugate $g * h * g^{-1}$ also belongs to H , since $g * h * g^{-1} \in gHg^{-1}$ and $(gH)g^{-1} = (Hg)g^{-1} = H(g * g^{-1}) = H$. So H is a union of conjugacy classes. $\square$

This proposition provides the answer we were looking for: the subgroups whose corresponding left and right cosets gH and Hg coincide, are exactly those subgroups which happen to be unions of conjugacy classes. In the case of an abelian group, we saw that the conjugacy classes are just sets consisting of individual elements. So any subgroup of an abelian group is therefore a union of conjugacy classes, and by Proposition 3.10 its left and right cosets must be the same. (In addition to the more obvious reason that all the elements commute.)

Equivalently, the left and right cosets of H in G coincide if and only if $g * h * g^{-1} \in H$ for all $g \in G$ and $h \in H$. We give a subgroup of this type a special name:

Definition 3.11 A subgroup $H < G$ is said to be a **normal subgroup** of G or **normal** in G if its left and right cosets coincide, in the sense that $gH = Hg$ for all $g \in G$.

Equivalently, by Proposition 3.10, H is normal in G if it is a union of conjugacy classes of elements of G . We denote normal subgroups by $H \triangleleft G$ or $H \triangleleft G$.

In the examples earlier on, we noticed that $R_3 \triangleleft D_3$ and $A_4 \triangleleft S_4$. In both of these cases, the subgroups have index 2 in the larger group: $|D_3 : R_3| = |S_4 : A_4| = 2$. It is certainly not the case that a normal subgroup must have index 2, since $3\mathbb{Z} \triangleleft \mathbb{Z}$ and $|\mathbb{Z} : 3\mathbb{Z}| = 3$. The converse, however, is true in general, as the following argument shows:

Proposition 3.12 Let $H < G$ be an index-2 subgroup of G . Then H is normal in G .

Proof Suppose that $|G : H| = 2$. Then there are only two distinct right cosets: H itself, and the complement $G \setminus H$. Similarly, there are only two distinct right cosets: H and $G \setminus H$ again. Clearly, each left coset is then equal to the corresponding right coset, and hence $H \triangleleft G$. $\square$

Example 3.13 For some geometric object X , the direct isometry group $\text{Isom}^+(X)$ is a normal subgroup of the full isometry group $\text{Isom}(X)$. This follows from the observation that

$$|\text{Isom}(X) : \text{Isom}^+(X)| = 2.$$

Example 3.14 $A_n \triangleleft S_n$ since $|S_n : A_n| = 2$ for any $n > 1$. This generalises Example 3.1.

Example 3.15 $SL_n(\mathbb{R}) \triangleleft GL_n(\mathbb{R})$ since every conjugate of a matrix with determinant 1 also has determinant 1, by the multiplicative property of matrix determinants. Suppose $A \in SL_n(\mathbb{R})$ (so $\det A = 1$) and that $B \in GL_n(\mathbb{R})$ with $\det B = d \in \mathbb{R}$. Then $\det(B^{-1}) = \frac{1}{d}$ and so $\det(BAB^{-1}) = \det(B) \det(A) \det(B^{-1}) = \frac{1}{d} \cdot 1 \cdot d = 1$. Hence $SL_n(\mathbb{R}) \triangleleft GL_n(\mathbb{R})$.

By a similar argument, $SO_n(\mathbb{R}) \triangleleft O_n(\mathbb{R})$ and $SU_n \triangleleft U_n$.

For the next two examples, we consider commutativity of individual elements. Recall that two group elements $g, h \in G$ **commute** if $g * h = h * g$. A group G is abelian if every element commutes with every other element. For any (abelian or nonabelian) group G , the identity element e commutes with every other element. It's often useful to consider which elements commute with which other elements.

In the dihedral group D_3 , for example, we see that every element of the rotation subgroup R_3 commutes with every other element of R_3 . But $r * m_1 = m_3 \neq m_2 = m_1 * r$, so r and m_1 don't commute with each other. The identity element e obviously commutes with everything, but nothing else does.

In the dihedral group D_4 , the 180° rotation r^2 commutes with every other element of the group, as does the identity element e , but nothing else does. We generalise this to obtain the following definition:

Definition 3.16 The **centre** $Z(G)$ of a group G is the subset of elements which commute with every element of G . That is,

$$Z(G) = \{g \in G : g * h = h * g \text{ for all } h \in G\}.$$

The centre of a group isn't just a subset, it's actually a normal subgroup, as the following proposition shows. (We'll make use of this fact later on when we study quotient groups.)

Proposition 3.17 *The centre $Z(G)$ of a group G is a normal subgroup.*

Proof First we show that $Z(G)$ is a subgroup of G . Clearly $e \in Z(G)$, since $e * g = g * e$ for all $g \in G$. The centre is closed under the induced operation since if $g, h \in Z(G)$, it follows that $g * h \in Z(G)$ too. That is,

$$(g * h) * k = g * (h * k) = g * (k * h) = (g * k) * h = (k * g) * h = k * (g * h)$$

for any $k \in G$. The second and fourth of these equalities follow since $g, h \in Z(G)$, and the rest are just consequences of associativity.

Also, if $g \in Z(G)$ then $g^{-1} \in Z(G)$, since

$$g^{-1} * h = (h^{-1} * g)^{-1} = (g * h^{-1})^{-1} = h * g^{-1}.$$

Hence $Z(G)$ is a subgroup of G . To show that it's a normal subgroup of G we need to confirm that for any $g \in Z(G)$ and $h \in G$, the conjugate $h * g * h^{-1}$ is also in $Z(G)$. Let $g \in Z(G)$ and $h \in G$. Then $h * g = g * h$ if and only if $g = h * g * h^{-1}$, so $g \in Z(G)$ exactly when g is conjugate only to itself, and hence its conjugacy class is the singleton set $\{g\}$. Hence $Z(G)$ is a union of (singleton) conjugacy classes, and is thus a normal subgroup of G . $\square$

This result is certainly consistent with what we know so far. If G is abelian then clearly $Z(G) = G$, since every element of G commutes with every other element of G . As noted earlier, the conjugacy classes of an abelian group G are the singleton sets $\{g\}$ for all $g \in G$. And the proposition above tells us that $Z(G)$ is the union of all the singleton conjugacy classes of G , which in this case is all of G .

Look at the conjugacy classes of D_3 again. These are $\{e\}$, $\{r, r^2\}$ and $\{m_1, m_2, m_3\}$. The only one of these which contains only a single element is $\{e\}$, so the centre $Z(D_3)$ is trivial, as noted earlier.

The conjugacy classes of D_4 are $\{e\}$, $\{r^2\}$, $\{r, r^3\}$ and $\{m_1, m_2, m_3, m_4\}$, and hence $Z(D_4) = \{e\} \cup \{r^2\} = \{e, r^2\}$.

Going back to D_3 again, we noticed that the identity e commuted with everything, and that r and r^2 commute with each other, but no other element commutes with anything apart from itself and the identity. That is, $e * g = g * e$ and $r * r^2 = r^2 * r$, but $g * h \neq h * g$ for any other $g, h \in D_3$. Another way to look at this is to take $g * h = h * g$ and rearrange it to give $g * h * g^{-1} * h^{-1} = e$. So g and h commute exactly when $g * h * g^{-1} * h^{-1} = e$; in other words the product $g * h * g^{-1} * h^{-1}$ measures how badly g and h fail to commute.

Definition 3.18 For any two elements $g, h \in G$ of a group G , we define the **commutator** $[g, h] := g * h * g^{-1} * h^{-1}$.

Table 3.2 lists all of the commutators in D_3 : the cell on the g th row and h th column contains the commutator $[g, h]$.

We can see that in D_3 the commutators $[g, h]$ are either the identity e (in which case the elements g and h commute) or a clockwise or anticlockwise rotation.

This tells us that, as we already knew, not all elements of D_3 commute with each other. But it also gives us some idea of how badly they fail to commute; in other words, the commutators measure how nonabelian the group D_3 is. In this case, they tell us that, at worst, two elements of D_3 fail to commute by a rotation of some sort.

Another way of looking at this is to say that D_3 would be abelian if we ignored rotations in some way. More precisely, if we defined some sort of equivalence relation on D_3 , where we considered two elements

	e	r	r^2	m_1	m_2	m_3
e	e	e	e	e	e	e
r	e	e	e	r^2	r^2	r^2
r^2	e	e	e	r	r	r
m_1	e	r	r^2	e	r^2	r
m_2	e	r	r^2	r	e	r^2
m_3	e	r	r^2	r^2	r	e

Table 3.2: Commutators in D_3

to be equivalent if they differed only by a rotation, then we could form a new, smaller group consisting of those equivalence classes. And if we did so, that group would be abelian. We'll come back to this idea in the next section when we study quotient groups.

Looking at the set of commutators in D_3 , we find it's exactly equal to the rotation subgroup $R_3 = \{e, r, r^2\}$. Is this a coincidence? Or is it always the case that the commutators in a group G form a subgroup?

As it happens, yes, in general it is slightly coincidental: the commutators of a group G don't, in general, form a subgroup of G .¹

We can fix this by slightly modifying our approach. The subset of commutators doesn't always form a subgroup, since the product of two commutators needn't itself be a commutator. However, if we consider the subgroup *generated by* the commutators (that is, the subgroup consisting of all possible commutators, their products, the products of those products, and so on) then everything works fine.

Definition 3.19 Given a group G , let $[G, G]$ denote the **commutator subgroup** of G , namely the subgroup generated by all possible commutators $[g, h] = g * h * g^{-1} * h^{-1}$ for $g, h \in G$. This is sometimes called the **derived group** of G . Some writers use the notation G' or $G^{(1)}$ instead of $[G, G]$.

¹ The smallest groups in which the set of commutators is not closed under the induced multiplication operation are of order 96. More concretely, Joseph Rotman implemented a computer search which found exactly two such nonisomorphic groups of order 96. One is a semidirect product $(\mathbb{Z}_2 \times \mathbb{Z}_2 \times Q_8) \rtimes \mathbb{Z}_3$ and the other is a semidirect product $((\mathbb{Z}_2 \times \mathbb{Z}_4) \rtimes \mathbb{Z}_4) \rtimes \mathbb{Z}_3$. We'll find out about semidirect products later.

	e	r	r^2	r^3	m_1	m_2	m_3	m_4
e	e	e	e	e	e	e	e	e
r	e	e	e	e	r^2	e	r^2	e
r^2	e	e	e	e	e	e	e	e
r^3	e	e	e	e	r^2	e	r^2	e
m_1	e	r^2	e	r^2	e	e	e	e
m_2	e	e	e	e	e	e	e	e
m_3	e	r^2	e	r^2	e	e	e	e
m_4	e	e	e	e	e	e	e	e

Table 3.3: Commutators in D_4

We've seen that the commutator subgroup $[D_3, D_3]$ is equal to the rotation subgroup $R_3 = \{e, r, r^2\}$. Is this true for other dihedral groups? That is, does $[D_n, D_n] = R_n$ for all n ? Table 3.3 lists the commutators in D_4 .

So $[D_4, D_4] = \{e, r^2\}$, which isn't the rotation subgroup $R_4 = \{e, r, r^2, r^3\}$. One of the exercises at the end of this chapter asks you to take this further and examine $[D_n, D_n]$ in greater generality.

Another important class of nonabelian groups we've met are the symmetric groups S_n . The following proposition characterises their commutator subgroups:

Proposition 3.20 The commutator subgroup $[S_n, S_n]$ is the alternating group A_n for $n \geq 3$.

Proof By the discussion following Definition 1.62, the commutator $[\sigma, \tau]$ of any two permutations $\sigma, \tau \in S_n$ must be even, since $[\sigma, \tau] = \tau^{-1}\sigma^{-1}\tau\sigma$ is a product of four permutations. If σ and τ are both odd, then $\tau\sigma$ and $\tau^{-1}\sigma^{-1}$ must both be even and hence $[\sigma, \tau]$ must also be even; if σ and τ are both even then $[\sigma, \tau]$ must also be even; and if (without loss of generality) σ is odd and τ is even, then both $\tau\sigma$ and $\tau^{-1}\sigma^{-1}$ must be odd, and their product $[\sigma, \tau]$ must be even.

Proposition 1.64 tells us that any even permutation (and hence any

element of A_n) can be written as a product of 3-cycles. Any 3-cycle $(a\ b\ c)$ can be decomposed as a product $(a\ b)(a\ c)(a\ b)(a\ c)$, which (since transpositions are self-inverse) is equal to $(a\ b)^{-1}(a\ c)^{-1}(a\ b)(a\ c) = [(a\ c), (a\ b)]$.

Hence all 3-cycles in S_n are commutators, and since A_n is generated by the 3-cycles in S_n , it is generated by commutators. Furthermore, there are no commutators in S_n which are not in A_n , since as noted above, all commutators have even parity. Therefore $[S_n, S_n] = A_n$. $\square$

The commutator subgroup $[G, G]$, like the centre $Z(G)$, provides a measure of how abelian a group G is. We saw earlier that $Z(G) = G$ exactly when G is abelian, and clearly $[G, G]$ is trivial exactly when G is abelian too. The centre turned out to be a normal subgroup of G , so what about the commutator subgroup?

Proposition 3.21 *The commutator subgroup $[G, G]$ of a group G is normal in G .*

Proof By the remark preceding Definition 3.11, we just need to show that the conjugate of a commutator (or a product of commutators) is also a commutator (or a product of commutators).

Suppose that $[g, h] = g * h * g^{-1} * h^{-1} \in [G, G]$. Then if $k \in G$ we have

$$\begin{aligned} k * [g, h] * k^{-1} &= k * (g * h * g^{-1} * h^{-1}) * k^{-1} \\ &= (k * g * k^{-1}) * (k * h * k^{-1}) * (k * g^{-1} * k^{-1}) * (k * h^{-1} * k^{-1}) \\ &= (k * g * k^{-1}) * (k * h * k^{-1}) * (k * g * k^{-1})^{-1} * (k * h * k^{-1})^{-1} \\ &= [k * g * k^{-1}, k * h * k^{-1}] \in [G, G]. \end{aligned}$$

So the conjugate of a commutator is also a commutator. But not every element of $[G, G]$ need be a commutator, so we have to check that the conjugate of a product of commutators is also a product of commutators.

Suppose that $x_1, \dots, x_n$ are commutators in $[G, G]$. Then $x_1 * \dots * x_n$ is also an element of $[G, G]$, and

$$k * x_1 * \dots * x_n * k^{-1} = (k * x_1 * k^{-1}) * \dots * (k * x_n * k^{-1}) \in [G, G].$$

So the conjugate of a product of commutators is also a product of commutators, and hence $[G, G]$ is a normal subgroup of G . $\square$

We will return to the commutator subgroup in the next section, where we will use it to form the **abelianisation** of a group G . But now we will use the commutator to prove a useful fact about normal subgroups.

Proposition 3.22 *Let H and K be normal subgroups of a group G such that $H \cap K = \{e\}$. Then the elements of H commute with the elements of K ; that is, $hk = kh$ for any $h \in H$ and $k \in K$.*

Proof Let $h \in H$ and $k \in K$ be arbitrary elements of the respective subgroups, and consider the commutator $[h, k] = hkh^{-1}k^{-1}$. Writing this as $h(kh^{-1}k^{-1})$, the parenthesised expression must lie in H because H is normal and hence closed under conjugation, and so $[h, k] \in H$. Similarly, writing $[h, k]$ as $(hkh^{-1})k$ we find that $[h, k] \in K$. Thus $[h, k] \in H \cap K$, and therefore $[h, k] = hkh^{-1}k^{-1} = e$, which gives $hk = kh$ as claimed. $\square$

This gives us a useful variation on Proposition 2.12, concerning the internal direct product of H and K :

Corollary 3.23 *Let H and K be normal subgroups of a group G such that $H \cap K = \{e\}$ and $HK = G$. Then $G \cong H \times K$.*

Proof By Proposition 3.22 every element of H commutes with every element of K , so all the hypotheses for Proposition 2.12 are satisfied, and hence $G \cong H \times K$. $\square$

All parts should go together without forcing. You must remember that the parts you are reassembling were disassembled by you. Therefore, if you can't get them together again, there must be a reason. By all means, do not use a hammer.

— IBM maintenance manual (c.1925)

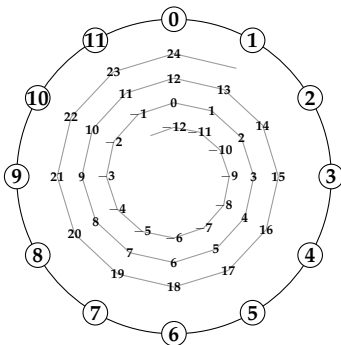


Figure 3.1: Wrapping $\mathbb{Z}$ into $\mathbb{Z}_{12}$.

3.2 Quotient groups

AS WE SAW in Corollary 2.28, given any subgroup $H < G$, we can neatly split G into a collection of subsets (called **cosets**), all of which are the same size. There are two different ways of doing this, depending on whether we look at the left or right cosets; in the last section we examined the circumstances in which these two ways might give the same result. In this section we investigate how these pieces fit together.

We start by going back to some of the earliest examples we considered. One way of defining the finite cyclic group $\mathbb{Z}_n$ is by taking the group $\mathbb{Z}$ of integers and defining the ‘congruence modulo n ’ equivalence relation $\equiv_n$ on it: $a \equiv_n b$ if $n \mid (b - a)$. We represented this schematically as a kind of clock face, but we can also think of it as wrapping the infinitely-long line of integers round in a circle so that integers with the same residue or remainder modulo n all line up: see Figure 3.1.

More formally, the equivalence relation $\equiv_n$ partitions $\mathbb{Z}$ into n equivalence classes $[0], \dots, [n-1]$. An element of $[a]$ plus an element of $[b]$ lies in $[a+b]$, so we can take the set $\{[0], \dots, [n-1]\}$ and define a new addition operation on this set such that $[a] + [b] = [a+b]$. This operation is obviously commutative and associative; furthermore the class $[0]$ acts as an identity since $[0] + [a] = [a] + [0] = [a]$, and for each class $[a]$, the class $[n-a]$ acts as an additive inverse since $[a] + [n-a] = [a+n-a] = [n] = [0]$. So these equivalence classes form a group, which is isomorphic to $\mathbb{Z}_n$.

Look again at Figure 3.1: all of the integers

$$\dots, -2n, -n, 0, n, 2n, \dots$$

are congruent to 0 and hence

$$[0] = \{\dots, -2n, -n, 0, n, 2n, \dots\} = n\mathbb{Z}.$$

Similarly,

$$[1] = \{\dots, 1-2n, 1-n, 1, 1+n, 1+2n, \dots\}$$

is the coset $1 + n\mathbb{Z}$ in $\mathbb{Z}$, and in fact $[a] = a + n\mathbb{Z}$ for all $0 \leq a < n$.

So it's starting to look like we can construct $\mathbb{Z}_n$ by taking the cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ and defining a suitable binary operation on them.

What we've done here is to take $\mathbb{Z}$ and factor by (or 'mod out by') congruence modulo n , considering two elements of $\mathbb{Z}$ to be 'the same' if they have the same remainder when divided by n . In effect, we're taking all of the integers which are congruent to each other, gluing them all together into n large clumps labelled $[0], \dots, [n-1]$, and then using these clumps as the elements of a new group, which turns out to be isomorphic to $\mathbb{Z}_n$.

Some questions arise at this point. Firstly, we did this with the subgroup $n\mathbb{Z} < \mathbb{Z}$. Will any subgroup work? Is the new operation, defined on the cosets of $n\mathbb{Z}$, well-defined? And how do we generalise this construction?

Let's look at another example. The dihedral group D_3 , the symmetry group of the equilateral triangle, has a three-element subgroup R_3 . Earlier, we defined a congruence relation $\equiv_n$ on $\mathbb{Z}$ such that $a \equiv_n b$ if $(a-b)$ is an element of the subgroup $n\mathbb{Z} < \mathbb{Z}$. The analogous construction for $R_3 < D_3$ is to define an equivalence relation $\sim$ such that $g \sim h$ if $g * h^{-1} \in R_3$. If we do this, D_3 splits neatly into two equivalence classes, which happen to be the left (and, since R_3 is normal in D_3 , right) cosets of R_3 :

$$[e] = \{e, r, r^2\} = R_3 \quad [m_1] = \{m_1, m_2, m_3\} = m_1 R_3 = R_3 m_1$$

We can immediately see from the multiplication table of D_3 that any two elements of $[e]$ combine to give another element of $[e]$, any two elements of $[m_1]$ combine to give an element of $[e]$, and any element of $[e]$ combined with any element of $[m_1]$ gives an element of $[m_1]$.

What this means is that the multiplication operation in D_3 lines up correctly with the way D_3 is partitioned into cosets of the subgroup R_3 . Just as we did with the congruence relation $\equiv_n$ defined on $\mathbb{Z}$, we can form a new group by treating the equivalence classes $[e]$ and $[m_1]$ as if they were single elements in their own right. The multiplication operation we get if we do this is shown in Table 3.4 (see also Table 3.5).

*	$[e]$	$[m_1]$
$[e]$	$[e]$	$[m_1]$
$[m_1]$	$[m_1]$	$[e]$

Table 3.4: Multiplication table for cosets of R_3 in D_3

*	e	r	r^2	m_1	m_2	m_3
e	e	r	r^2	m_1	m_2	m_3
r	r	r^2	e	m_3	m_1	m_2
r^2	r^2	e	r	m_2	m_3	m_1
m_1	m_1	m_2	m_3	e	r	r^2
m_2	m_2	m_3	m_1	r^2	e	r
m_3	m_3	m_1	m_2	r	r^2	e

Table 3.5: Cosets of R_3 in D_3

This is obviously just the multiplication table for $\mathbb{Z}_2$ in a very superficial disguise. So our new group $\{[e], [m_1]\}$ with multiplication operation derived from the original multiplication operation $*$ in D_3 is isomorphic to the two-element finite cyclic group $\mathbb{Z}_2$.

Geometrically, the elements of $[e]$ are the orientation-preserving isometries of the equilateral triangle Δ , while the elements of $[m_1]$ are exactly those isometries of Δ which reverse the orientation. So what this tells us is that two direct isometries preserve orientation, and two opposite isometries do as well, but a direct isometry combined with an opposite isometry reverses orientation. The way direct and opposite isometries interact is the same as the way the elements 0 and 1 interact in the group $\mathbb{Z}_2$.

Let's introduce the symbol $/$ to denote this (as yet insufficiently precisely defined) operation on groups and their subgroups. So the previous two examples tell us that

$$\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n \quad \text{and} \quad D_3/R_3 \cong \mathbb{Z}_2.$$

To decide whether this construction works for any subgroup H of any group G , we'll try another example. Let M denote the subgroup $\{e, m_1\}$ of D_3 . This has three left cosets

$$M = \{e, m_1\} \quad rM = \{r, m_3\} \quad r^2M = \{r^2, m_2\}$$

and three right cosets

$$M = \{e, m_1\} \quad Mr = \{r, m_2\} \quad Mr^2 = \{r^2, m_3\}$$

from which we can see immediately that M is, unlike R_3 , not a normal subgroup of D_3 . In particular, this means that whereas with the previous examples we could cheerfully ignore whether the resulting groups consisted of left or right cosets, we have to be more careful in this case. For the moment, let's consider the set of left cosets of M in D_3 and see how they interact. Following on from the previous examples, given two elements $g, h \in D_3$, we require that the product of any element of the left coset gM with any element of the left coset hM be an element of the left coset $(g * h)M$. Unfortunately, this doesn't always work. For example,

$$(eM)(rM) = \{g * h : g \in M, h \in rM\} = \{r, r^2, m_2, m_3\}$$

which not only fails to be the left coset rM , but isn't even one of the other two left cosets either. Looking at the multiplication table for D_3 again, this time rearranging the elements into their left coset groups, we can see (in Table 3.6) that instead of a nice, neat structure like the one we got for D_3/R_3 , the whole thing has fragmented in a decidedly inelegant manner.

*	e	m_1	r	m_3	r^2	m_2
e	e	m_1	r	m_3	r^2	m_2
m_1	m_1	e	m_2	r^2	m_3	r
r	r	m_3	r^2	m_2	e	m_1
m_3	m_3	r	m_1	e	m_2	r^2
r^2	r^2	m_2	e	m_1	r	m_3
m_2	m_2	r^2	m_3	r	m_1	e

Table 3.6: Left cosets of $\{e, m_1\}$ in D_3

What's going wrong here? What precisely is stopping the left cosets of M in D_3 from behaving like elements of a group?

We want to be able to define a group structure on the set of left cosets of M such that

$$(gM)(hM) = (g * h)M$$

for all $g, h \in D_3$. But the problem is that the set $\{M, rM, r^2M\}$ isn't closed under this multiplication operation, because the multiplication operation in D_3 doesn't line up neatly with the left cosets of M in D_3 .

It works in some cases. For example, $(rM)(eM) = \{r, m_3\} = rM$ as required. But if we try $(eM)(rM)$ we find that it doesn't. In particular,

$$\begin{aligned} m_1 * r &= (e * m_1) * (r * e) \\ &= e * (r * r^{-1}) * m_1 * r * e \\ &= (e * r) * (r^{-1} * m_1 * r) * e. \end{aligned}$$

In the last line, the first parenthesised expression $(e * r)$ is equal to r , as we want it to be. But in order for this whole expression to be an element of rM , we also need the second parenthesised expression $(r^{-1} * m_1 * r)$ to be an element of M , which it isn't since

$$r^{-1} * m_1 * r = r^2 * m_1 * r = m_3.$$

Notice that $(r^{-1} * m_1 * r)$ is a conjugate of an element of M . In fact, if we tried any other combination of cosets of M (rM and r^2M , for example) we'd end up with a similar requirement: that a particular conjugate of an element of M be itself an element of M .

In other words, the reason this particular example doesn't work is because M isn't closed under conjugation. In the last section, we introduced a special class of subgroups which *are* closed under conjugation, namely normal subgroups. So the reason we can't form a group from the left cosets of M is because M isn't a normal subgroup of D_3 . The rotation subgroup *is* normal in D_3 , and the subgroup $n\mathbb{Z}$ is normal in $\mathbb{Z}$, which is why this construction worked properly in those cases.

The following proposition states this more generally:

Proposition 3.24 *If H is a normal subgroup of G , the set of all left cosets of H in G forms a group under the multiplication operation*

$$(gH)(kH) = (g * k)H$$

for all $g, k \in G$.

Proof Any element of $(gH)(kH)$ has the form $g * h_1 * k * h_2$, for some $h_1, h_2 \in H$. We can rewrite this as

$$g * h_1 * k * h_2 = g * (k * k^{-1}) * h_1 * k * h_2 = (g * k) * (k^{-1} * h_1 * k) * h_2.$$

This is an element of $(g * k)H$ if and only if the conjugate $(k^{-1} * h_1 * k)$

is an element of H . But since H is a normal subgroup of G , it's closed under conjugation (by Proposition 3.10) and hence $(k^{-1} * h_1 * k)$ must also lie in H . So the set of left cosets of H in G is closed under the multiplication operation defined above, as required.

In order to show that the set of left cosets forms a group, we must check each of the group axioms. The new multiplication operation is clearly associative, because the operation $*$ is. The coset eH acts as the identity element since

$$(eH) * (gH) = gH = (gH) * (eH)$$

for any $g \in G$. Finally, the coset $g^{-1}H$ is the inverse of the coset gH since

$$(g^{-1}H)(gH) = eH = (gH)(g^{-1}H)$$

for any $g \in G$. □

This proposition and discussion leads to the following definition:

Definition 3.25 Let H be a normal subgroup of a group G . Then the **quotient group** or **factor group** G/H is the group consisting of the left cosets of H in G together with the multiplication operation

$$(gH)(kH) = (g * k)H$$

for any $g, k \in G$.

To better understand this concept, we'll look at a few examples.

Example 3.26 The cyclic group $\mathbb{Z}_4$ has a subgroup $\langle 2 \rangle = \{0, 2\}$ isomorphic to $\mathbb{Z}_2$. Its cosets are

$$\langle 2 \rangle = 0 + \langle 2 \rangle = \{0, 2\} \quad 1 + \langle 2 \rangle = \{1, 3\}$$

The subgroup $\langle 2 \rangle$ is normal since $\mathbb{Z}_4$ is abelian, so we can go ahead and construct the quotient group $\mathbb{Z}_4 / \langle 2 \rangle = \{\langle 2 \rangle, 1 + \langle 2 \rangle\}$ which has two elements and must therefore be isomorphic to $\mathbb{Z}_2$.

We can also see this from a rearranged version of the multiplication table for $\mathbb{Z}_4$, shown in Table 3.7.

+	0	2	1	3
0	0	2	1	3
2	2	0	3	1
1	1	3	2	0
3	3	1	0	2

 $\cong$

+	0	1
0	0	1
1	1	0

Table 3.7: $\mathbb{Z}_4 / \langle 2 \rangle \cong \mathbb{Z}_2$

*	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

 $\cong$

+	0	1
0	0	1
1	1	0

Table 3.8: $V_4 / \{e, a\} \cong \mathbb{Z}_2$

Example 3.27 The Klein 4-group $V_4 = \{e, a, b, c\} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$ has three proper, nontrivial subgroups, each of which is isomorphic to $\mathbb{Z}_2$. Consider the subgroup $A = \{e, a\}$; again this is normal since V_4 is abelian. Its cosets are A and $bA = cA = \{b, c\}$ hence

$$V_4 / A = \{A, bA\} \cong \mathbb{Z}_2.$$

This again can be seen from the multiplication table in Table 3.8.

The above example could be written, with only a slightly ambiguous abuse of notation, as $(\mathbb{Z}_2 \oplus \mathbb{Z}_2) / \mathbb{Z}_2 \cong \mathbb{Z}_2$.

Let's try another example like this.

Example 3.28 Let $G = \mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3$ (by Proposition 1.32). Now consider the subgroup $H = \langle 3 \rangle \cong \mathbb{Z}_2 \oplus \{0\}$. It has the following cosets:

$$0+H = \{0, 3\} \cong \{(0, 0), (1, 0)\} = (0, 0)+H$$

$$1+H = \{1, 4\} \cong \{(0, 1), (1, 1)\} = (0, 1)+H$$

$$2+H = \{2, 5\} \cong \{(0, 2), (1, 2)\} = (0, 2)+H$$

Furthermore, H is normal in G , not least because G is abelian. So we can form the quotient group

$$G/H = \{H, 1+H, 2+H\} \cong \{(0, 0)+H, (0, 1)+H, (0, 2)+H\}$$

which must be isomorphic to $\mathbb{Z}_3$, the only group with three elements. Again, the multiplication table, suitably rearranged, makes this clear: see Table 3.9

+	0	3	1	4	2	5
0	0	3	1	4	2	5
3	3	0	4	1	5	2
1	1	4	2	5	3	0
4	4	1	5	2	0	3
2	2	5	3	0	4	1
5	5	2	0	3	1	4

 $\cong$

+	0	1	2
0	0	0	1
1	1	1	2
2	2	2	0

Table 3.9: $\mathbb{Z}_6/\langle 3 \rangle \cong \mathbb{Z}_3$

Is this true in general? That is, given a direct product $G \times H$ of groups, is it true that $\{e\} \times H \cong H$ is a normal subgroup of $G \times H$, and if so, that $(G \times H)/(\{e\} \times H) \cong G$? Yes, as the next proposition confirms.

Proposition 3.29 *The direct product $G \times H$ of any two groups G and H has a normal subgroup isomorphic to H . Factoring out by this subgroup yields a group isomorphic to G .*

Proof The direct product $G \times H$ is the set $\{(g, h) : g \in G, h \in H\}$ of ordered pairs of elements of G and H . The subset $\{e\} \times H = \{(e, h) : h \in H\}$ may be seen to be a subgroup of $G \times H$; furthermore it is isomorphic to H via the map $f: \{e\} \times H \rightarrow H$ given by $f: (e, h) \mapsto h$.

To see that this subgroup is normal in $G \times H$, choose any $g \in G$ and $h, k \in H$. Then the conjugate

$$(g, k) * (e, h) * (g, k)^{-1} = (g * e * g^{-1}, k * h * k^{-1}) = (e, k * h * k^{-1})$$

is an element of $\{e\} \times H$, and hence $\{e\} \times H$ is closed under conjugation and thus normal.

The cosets of $\{e\} \times H$ in $G \times H$ are of the form $(g, h)(\{e\} \times H) = \{g\} \times H$ and are thus indexed by elements of G . Therefore the quotient group $(G \times H)/(\{e\} \times H)$ is isomorphic to G . $\square$

What this tells us is that any direct product of a finite number of groups can be divided by any of its factors.

It might, therefore, be tempting to think that any group can be decomposed in this way. That is, that $(G/H) \times H \cong G$ for any normal subgroup $H < G$. But Examples 3.27 and 3.26 above show that this isn't the case: both $\mathbb{Z}_4$ and V_4 contain a subgroup isomorphic to $\mathbb{Z}_2$, but we know from Example 1.30 that $\mathbb{Z}_4 \not\cong V_4$. So not every group is isomorphic to a direct product of its various quotients.

In fact, we can construct $\mathbb{Z}_4$ from two copies of $\mathbb{Z}_2$, but we have to do it in a slightly more complicated way, using something called an **extension**. This is beyond the scope of the current discussion, but we'll come back to it later, in Section 7.5.

It's worth thinking about what the operation of factoring by a normal subgroup actually means, in addition to the purely algebraic definition in terms of cosets. Where the group in question has a particular geometric or other interpretation, what does the quotient operation mean in that context?

We've already said that factoring a group G by a normal subgroup H can be thought of as gluing together the elements of a coset into a single block and then looking at how they interact with each other under the group multiplication operation. In particular, though, the subgroup H itself collapses down to become the identity element in the quotient group G/H . In doing so, everything that the subgroup H represented, measured or related to in G gets discarded or ignored in the passage to G/H .

Looking at our first example $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$, what's going on here is that in $\mathbb{Z}$ we can uniquely express any integer m in the form $m = qd + r$, where $0 \leq r < n$ and $q \in \mathbb{Z}$. That is, we can consider any integer m as being formed from an element of $n\mathbb{Z}$, a multiple of n , together with an extra bit, the remainder r . In passing to the quotient group $\mathbb{Z}/n\mathbb{Z}$, we're essentially ignoring the multiple of n bit and only keeping the remainder r .

Similarly, when we factored the dihedral group D_3 by the rotation subgroup R_3 , we were ignoring any rotational component of an element in D_3 , that is, a symmetry of the equilateral triangle, and only looking at what was left: the reflection component. Geometrically, we can think of this as taking the original equilateral triangle, gluing together (or **identifying**) all the points which differ by a rotation, and thereby obtaining a new shape (see Figure 3.2). This new shape, similar to a kite, clearly only has a single line of symmetry, which is consistent with the quotient group D_3/R_3 being isomorphic to $\mathbb{Z}_2$.

As noted in the last section, every group has at least two normal subgroups: itself and the trivial subgroup $\{e\}$. Factoring by these two subgroups doesn't, however, yield anything particularly illuminating:

$$G/\{e\} \cong G \quad \text{and} \quad G/G \cong \{e\}$$

Convince yourself that both of these statements are true.

Two other subgroups that we met in the last section are the centre $Z(G)$ and the commutator subgroup $[G, G]$, both of which gave some measure of the commutativity (or not) of the group operation. These are normal subgroups, so we can use them to form quotient groups.

Recall that the centre $Z(G)$ consists of those elements of G which commute with all other elements of G . So $Z(G) = G$ if and only if G is abelian. If G is nonabelian then $Z(G)$ will be a proper normal

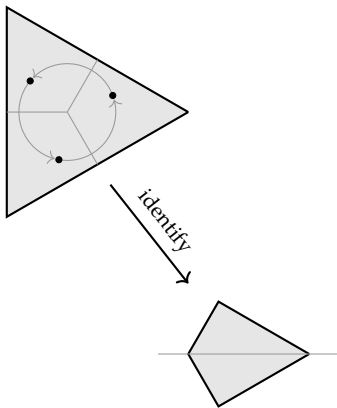


Figure 3.2: Identification of points in a triangle that differ by a $\frac{1}{3}$ -rotation

subgroup of G . Factoring by $Z(G)$ to get the quotient group $G/Z(G)$ can be regarded as ignoring the nicely commuting elements of G ; what's left gives some measure of how nonabelian G is. If $G/Z(G)$ is trivial then G must be abelian, while if $G/Z(G)$ is nontrivial then G is nonabelian to a greater or lesser extent.

Example 3.30 The centre $Z(D_3)$ is trivial, so $D_3/Z(D_3) \cong D_3$. So in some vague sense, D_3 is about as nonabelian as we can reasonably expect it to be.

The centre $Z = Z(D_4) = \{e, r^2\}$ so D_4/Z has four elements, and is hence isomorphic to either $\mathbb{Z}_4$ or $V_4 = \mathbb{Z}_2 \oplus \mathbb{Z}_2$. To see which one, we need to look at the cosets of Z :

$$\begin{aligned} Z &= \{e, r^2\} & rZ &= \{r, r^3\} \\ m_1Z &= \{m_1, m_3\} & m_2Z &= \{m_2, m_4\} \end{aligned}$$

The quotient group D_4/Z is isomorphic to the Klein 4-group $V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. This can be seen from the coset multiplication table (Table 3.10) or by observing that the cosets rZ , m_1Z and m_2Z all have order 2 under the coset multiplication operation.

*	eZ	rZ	m_1Z	m_2Z
eZ	eZ	rZ	m_1Z	m_2Z
rZ	rZ	eZ	m_2Z	m_1Z
m_1Z	m_1Z	m_2Z	eZ	rZ
m_2Z	m_2Z	m_1Z	rZ	eZ

Table 3.10: Coset multiplication table for $D_4/Z(D_4) \cong V_4$

We'll come back to this later in Section 7.2 when we use the quotient $G/Z(G)$ to form the **upper central series** of a group, but for the moment just think of it as giving an indication of the nonabelianness of a group.

The commutator subgroup $[G, G]$ of a group G is also a measure (albeit a slightly different one) of the nonabelianness of G : if $[G, G]$ is trivial then G is abelian, and the more elements $[G, G]$ has, the more nonabelian G is. Proposition 3.21 confirms that $[G, G]$ is normal in G , so we can form the quotient group $G/[G, G]$. In doing this, we are killing off all of the noncommutativity in G . More precisely, we are making every commutator $g*h*g^{-1}*h^{-1}$ equal to the identity e , forcing every element of G to commute with every other element of G .

Proposition 3.31 *For any group G , the quotient $G/[G, G]$ is abelian. Furthermore, if $N \trianglelefteq G$ and G/N is abelian, then $[G, G] \leq N$.*

Proof Suppose that $a, b \in G$, and let $G' = [G, G]$. We want to show that the cosets aG' and bG' commute in the quotient group G/G' , and we do this by considering their commutator:

$$\begin{aligned} [aG', bG'] &= (aG')(bG')(aG')^{-1}(bG')^{-1} = \\ &= (aba^{-1}b^{-1})G' = [a, b]G' = G'. \end{aligned}$$

(The last step follows because $[a, b] \in G'$.) This means that the commutator of $[aG', bG']$ of any two cosets in G/G' is trivial, and hence G/G' is abelian.

To prove the second statement, we consider the commutator $[aN, bN]$ of any two cosets aN and bN in the quotient G/N :

$$[aN, bN] = (aN)(bN)(aN)^{-1}(bN)^{-1} = (aba^{-1}b^{-1})N = [a, b]N.$$

Since G/N is abelian, this commutator must be the identity element in G/N , namely N . So $N = [a, b]N$ for any $a, b \in G$. By Proposition 2.26 this means that $[a, b] \in N$ and hence $G' \leq N$. $\square$

This process, of factoring a group by its commutator subgroup, is called **abelianisation**, and we will often denote the quotient $G/[G, G]$ by G^{ab} ; it is the largest abelian quotient of G .

Example 3.32 The commutator subgroup $[D_3, D_3]$ is the rotation subgroup $R_3 = \{e, r, r^2\}$, so $D_3^{\text{ab}} = D_3/R_3 \cong \mathbb{Z}_2$.

The commutator subgroup $[D_4, D_4]$ is the two-element subgroup $\{e, r^2\}$, so (by the discussion in Example 3.30) we see that

$$D_4^{\text{ab}} = D_4/[D_4, D_4] \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2.$$

Example 3.33 The alternating group A_n is a normal subgroup of S_n ; as remarked earlier (in Example 3.14) it has index 2 and is hence normal (by Proposition 3.12) and so $S_n/A_n \cong \mathbb{Z}_2$. In fact, for $n \geq 3$, Proposition 3.20 tells us that $[S_n, S_n] = A_n$, so we've just shown that $S_n^{\text{ab}} \cong \mathbb{Z}_2$ for $n \geq 3$.

Example 3.34 Generalising the examples of D_3 and D_4 above, the direct isometry group $\text{Isom}^+(X)$ of some geometric object X is an index-2 normal subgroup of the full isometry group $\text{Isom}(X)$, so

$$\text{Isom}(X)/\text{Isom}^+(X) \cong \mathbb{Z}_2.$$

The next two examples consider quotient groups of infinite groups.

Example 3.35 The multiplicative group $\mathbb{R}_+^\times$ of positive real numbers is a subgroup of the multiplicative group $\mathbb{C}_*^\times$ of nonzero complex numbers. In particular, $\mathbb{R}_+^\times$ is normal in $\mathbb{C}_*^\times$ since $\mathbb{C}_*^\times$ is abelian. Writing any element of $\mathbb{C}_*^\times$ as $re^{i\theta}$ where $r > 0$ and $\theta \in [0, 2\pi)$ we can see that the subgroup $\mathbb{R}_+^\times$ is exactly those elements of $\mathbb{C}_*^\times$ with $\theta = 0$. So we can specify any coset $re^{i\theta}\mathbb{R}_+^\times$ as $e^{i\theta}\mathbb{R}_+^\times$; this is uniquely determined by just the argument θ .

Furthermore,

$$(e^{i\theta}\mathbb{R}_+^\times)(e^{i\phi}\mathbb{R}_+^\times) = (e^{i(\theta+\phi)})\mathbb{R}_+^\times,$$

so the quotient $\mathbb{C}_*^\times/\mathbb{R}_+^\times$ is isomorphic to an additive group $\mathbb{R}$ of real numbers. But we're only interested in the **principal value** of the argument; that is, the angle it represents in $[0, 2\pi)$. So $\mathbb{C}_*^\times/\mathbb{R}_+^\times$ is isomorphic to the additive group of real numbers modulo 2π .

Geometrically, we can regard $\mathbb{C}_*^\times / \mathbb{R}_+^\times$ as the group of rotations in the plane $\mathbb{R}^2$. That is, the matrix group

$$\left\{ \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} : \theta \in [0, 2\pi) \right\}.$$

This is exactly the special orthogonal group $SO_2(\mathbb{R})$. So $\mathbb{C}_*^\times / \mathbb{R}_+^\times \cong SO_2(\mathbb{R})$. This is also isomorphic to the unitary group $U_1 = \{[e^{i\theta}] : \theta \in [0, 2\pi)\}$.

Alternatively, we can view the cosets as infinite radii in the complex plane; when we factor by $\mathbb{R}_+^\times$, we collapse each of these radii onto its intersection point with the unit circle.

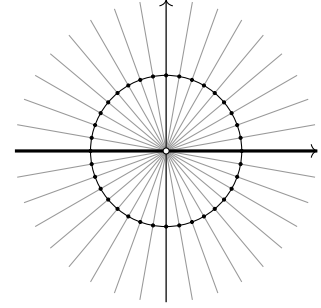


Figure 3.3: Cosets of $\mathbb{R}_+^\times$ in $\mathbb{C}_*^\times$

Example 3.36 The multiplicative group $\mathbb{C}_*^\times$ has another obvious subgroup consisting of the complex numbers of unit modulus:

$$C = \{z \in \mathbb{C}_*^\times : |z| = 1\} = \{e^{i\theta} : \theta \in [0, 2\pi)\} \cong U_1 \cong SO_2(\mathbb{R})$$

Again, C is normal in $\mathbb{C}_*^\times$ since the larger group is abelian, so we can form the quotient group $\mathbb{C}_*^\times / C$.

Since every nonzero complex number can be written in the form $re^{i\theta}$ where $r > 0$ and $\theta \in [0, 2\pi)$, we can express any coset $re^{i\theta}C = rC$. So a given coset of C is determined entirely by the modulus r , and hence we have

$$(rC)(sC) = (rs)C.$$

Thus $\mathbb{C}_*^\times / C \cong \mathbb{R}_+^\times$, the multiplicative group of positive real numbers. Geometrically, the cosets are concentric circles in the complex plane, and when we factor by C we are collapsing each circle down onto its intersection point with the positive real axis.

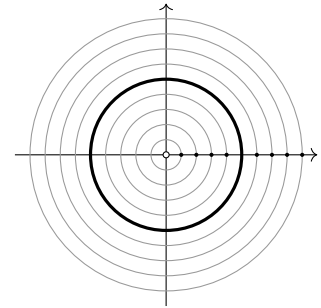


Figure 3.4: Cosets of C in $\mathbb{C}_*^\times$

We can take Example 3.35 a step further by replacing the nonzero complex numbers $\mathbb{C}_*^\times$ by the nonzero **quaternions** $\mathbb{H}_*^\times$.

Definition 3.37 A **quaternion** is a four-dimensional analogue of a (two-dimensional) complex number, an element of the form

$$z = a + bi + cj + dk$$

where $a, b, c, d \in \mathbb{R}$ and $ijk = i^2 = j^2 = k^2 = -1$. Together, they form an algebra $\mathbb{H}$, the third in a sequence $\mathbb{R}, \mathbb{C}, \mathbb{H}, \mathbb{O}$, formed by a process called the **Cayley–Dickson construction**. (The fourth algebra $\mathbb{O}$ is eight-dimensional and consists of objects called **octonions**.)

The algebra $\mathbb{H}$ is noncommutative, as the three special elements i, j and k **anticommute**:

$$\begin{array}{lll} ij = k & jk = i & ki = j \\ ji = -k & kj = -i & ik = -j \end{array}$$



Wikimedia Commons

The Irish mathematician and physicist Sir **William Rowan Hamilton** (1805–1865) displayed a strong mathematical talent from an early age, and was appointed Professor of Astronomy at the University of Dublin, while still a 21-year-old undergraduate. During his career he made many contributions to mechanics, optics and algebra.

He spent some years unsuccessfully trying to construct a three-dimensional analogue of the complex numbers until, while walking along the banks of the Royal Canal in Dublin on the afternoon of Monday 16 October 1843, he realised the solution: to construct instead a four-dimensional, noncommutative algebra. With his walking stick, he scratched the equations of quaternion multiplication on one of the stones of nearby Broom Bridge. In 1958, a commemorative plaque was unveiled by the then Taoiseach Éamon de Valera (1882–1975), also a mathematics graduate of the University of Dublin.



Geograph Project / © 2007 JP

(Readers interested in visiting the bridge may find it easiest to catch a local train to Broombridge Station.)

(The set $\{\pm 1, \pm i, \pm j, \pm k\}$, equipped with this multiplication operation, is the quaternion group Q_8 from Example 1.42.)

The subgroup $\mathbb{R}_+^\times$ is normal in $\mathbb{H}_*^\times$, and by a similar argument as in Example 3.35 we can show that $\mathbb{H}_*^\times / \mathbb{R}_+^\times \cong SU_2$, which happens to include as an index-2 (and hence normal) subgroup $SO_3(\mathbb{R})$, the group of rotations in three-dimensional space. This fact is particularly useful in computer graphics programming: quaternion multiplication turns out to be a computationally efficient way of calculating three-dimensional rotations.

In the case of the octonions, we not only have to give up commutativity of multiplication, but associativity as well. It is possible to take this sequence one step further, obtaining a sixteen-dimensional algebra variously called **hexadecanions** or **sedonions**, but we have to give up something else along the way: this new algebra is not only noncommutative and nonassociative, it also fails to be a **division algebra**, in the sense that there exist nonzero sedonions that nevertheless multiply to give zero. We'll learn about division algebras and division rings in Chapter 10.

We saw earlier that taking the quotient of a group G by its centre $Z(G)$ yields another group. If G was abelian to start with, then this quotient will simply be the trivial group $\{e\}$, but if G was nonabelian, then we'll get a nontrivial group (which itself might or might not be abelian). Applying this to the various families of matrix groups we met in Section 1.2 produces some more interesting examples of groups, some of which we'll investigate further in the next section.

Example 3.38 The centre $Z(GL_n(\mathbb{R}))$ of the general linear group $GL_n(\mathbb{R})$ consists of all nonzero scalar multiples kI of the identity matrix I . Factoring out by this group gives the **projective general linear group** $PGL_n(\mathbb{R})$. Applying the same process to the special linear group $SL_n(\mathbb{R})$ we obtain the **projective special linear group** $PSL_n(\mathbb{R})$.

In a similar way, the orthogonal group $O_n(\mathbb{R})$ gives rise to the **projective orthogonal group** $PO_n(\mathbb{R})$ and the special orthogonal group $SO_n(\mathbb{R})$ yields the **projective special orthogonal group** $PSO_n(\mathbb{R})$.

We can also construct the **projective unitary group** PU_n and the **projective special unitary group** PSU_n from U_n and SU_n , and the **projective symplectic group** $PSp_{2n}(\mathbb{R})$ from $Sp_{2n}(\mathbb{R})$.

In the next chapter we will return to this idea of factoring a group by its centre. In particular, Corollary 4.58, a consequence of the Normaliser–Centraliser Theorem,² gives us a concrete interpretation of this quotient.

² Theorem 4.57, page 124.

3.A Simple groups

WE LEARNED EARLIER that if p is prime, the cyclic group $\mathbb{Z}_p$ has no subgroups except for the trivial subgroup itself. This means $\mathbb{Z}_p$ is in some sense indivisible: it has no proper, nontrivial normal subgroups, so can't be factored to give a smaller nontrivial group.

Just as many properties of the multiplicative structure of the integers relate to the study of prime numbers, we can reasonably expect to understand some aspects of the classification of finite groups and their internal structure by studying those finite groups which can't be turned into smaller groups by factoring by some normal subgroup.

Definition 3.39 A group G is **simple** if it contains no proper, nontrivial normal subgroups. That is, it contains no normal subgroups other than G and $\{e\}$.

The classification of finite simple groups represents the end product of arguably the greatest collaborative project in the history of group theory. Earlier discussions yield the following two examples.

Example 3.40 The cyclic group $\mathbb{Z}_p$ is simple if and only if p is prime.

Example 3.41 The dihedral group D_n is not simple, since it contains a normal subgroup $R_n = \langle r \rangle$ generated by the rotations.

In fact, there are relatively few types of simple groups. The only other ones we've met so far, apart from the prime order cyclic groups $\mathbb{Z}_p$ are the alternating groups A_n for $n \neq 4$. The groups A_1 and A_2 are trivial (and hence simple). The group A_3 is isomorphic to $\mathbb{Z}_3$ and is hence simple too. The group A_4 , however, is not simple: it has a single proper, nontrivial normal subgroup

$$\{\iota, (12)(34), (13)(24), (14)(23)\} \cong V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2.$$

All the other alternating groups A_n for $n \geq 5$ are simple. To prove this, we first need the following lemma.

Lemma 3.42 Let $N \trianglelefteq A_n$, where $n \geq 5$. If N contains a 3-cycle, then $N = A_n$.

Proof Suppose N contains a 3-cycle $\alpha = (a_1 a_2 a_3)$. Let $\beta = (x y z)$ be another 3-cycle in A_n . Choose $\sigma \in S_n$ such that $\sigma(a_1) = x$, $\sigma(a_2) = y$ and $\sigma(a_3) = z$. Then $\sigma\alpha\sigma^{-1} = (x y z) = \beta$, because

$$x \xrightarrow{\sigma^{-1}} a_1 \xrightarrow{\alpha} a_2 \xrightarrow{\sigma} y,$$

$$y \xrightarrow{\sigma^{-1}} a_2 \xrightarrow{\alpha} a_3 \xrightarrow{\sigma} z,$$

$$\text{and } z \xrightarrow{\sigma^{-1}} a_3 \xrightarrow{\alpha} a_1 \xrightarrow{\sigma} x.$$

It would be of the greatest interest if a classification could be given of all simple groups with finitely many operations.

— Otto Hölder (1859–1937),
Die einfachen Gruppen in ersten und zweiten Hundert der Ordnungszahlen,
 Mathematische Annalen 40 (1892)
 55–88

Only three people have ever really understood the Schleswig–Holstein business: the Prince Consort, who is dead, a German professor, who is mad, and I, who have forgotten all about it.

— Henry John Temple, 3rd Viscount Palmerston (1784–1865)

(Remember that composition of permutations is written right-to-left, rather than left-to-right.) Since $n \geq 5$ there are at least two elements a_4 and a_5 not affected by α . Hence $\tau = (a_4 a_5)$ commutes with α and so

$$\beta = \sigma \alpha \sigma^{-1} = \sigma \tau \alpha \tau^{-1} \sigma^{-1} = (\sigma \tau) \alpha (\sigma \tau)^{-1}.$$

Either σ or $\sigma \tau$ belongs to A_n : either σ is an even permutation (and hence an element of A_n) or it's an odd permutation, in which case $\sigma \tau$ is an even permutation, since τ is a transposition disjoint from σ .

Hence α is conjugate to β in A_n , and so $\beta \in N$. Therefore N contains all 3-cycles in A_n and must therefore be the whole of A_n . $\square$

We can now prove that the alternating group A_n is simple for $n \geq 5$. The proof is a little involved, but essentially consists of applying Lemma 3.42 to five possible cases: an element of A_n (for $n \geq 5$) can

- (i) consist of a single 3-cycle,
- (ii) contain at least one cycle of length $m > 3$,
- (iii) consist of two or more disjoint 3-cycles together with zero or more disjoint transpositions,
- (iv) consist of a single 3-cycle and a product of disjoint transpositions, or
- (v) consist of an even number of disjoint transpositions.

Proposition 3.43 *The alternating group A_n is simple for $n \geq 5$.*

Proof Let $N \triangleleft A_n$ be nontrivial. We will show that N must be the whole of A_n . Consider an arbitrary element $\omega \in N$ and decompose it as a product of disjoint cycles. Then we have five cases to consider.

Case 1 $\omega = (a_1 a_2 a_3)$ is a 3-cycle. If so, then by Lemma 3.42 the subgroup N must consist of all the 3-cycles, and therefore be A_n itself.

Case 2 $\omega = \tau(a_1 a_2 \dots a_m)$, where $m > 3$ and τ is some product of disjoint cycles. In this case, the 3-cycle $\sigma = (a_1 a_2 a_3)$ commutes with all the factors of ω except for the m -cycle $(a_1 a_2 \dots a_m)$. So

$$\chi = \sigma \omega \sigma^{-1} = \tau \sigma (a_1 a_2 \dots a_m) \sigma^{-1}$$

belongs to N , and so does

$$\begin{aligned} \omega^{-1} \chi &= (a_1 a_2 \dots a_m)^{-1} \tau^{-1} \tau \sigma (a_1 a_2 \dots a_m) \sigma^{-1} \\ &= (a_m \dots a_2 a_1) (a_1 a_2 a_3) (a_1 a_2 \dots a_m) (a_3 a_2 a_1) \\ &= (a_1 a_3 a_m) \end{aligned}$$

which is a 3-cycle, and hence by Lemma 3.42 N must be A_n .

Case 3 $\omega = \tau(a_1 a_2 a_3)(a_4 a_5 a_6)$, where τ is a product of disjoint cycles. Let $\sigma = (a_2 a_3 a_4)$, and note that σ commutes with τ . Then

$$\begin{aligned} \omega^{-1} \sigma \omega \sigma^{-1} &= (a_6 a_5 a_4) (a_3 a_2 a_1) (a_2 a_3 a_4) (a_1 a_2 a_3) (a_4 a_5 a_6) (a_4 a_3 a_2) \\ &= (a_1 a_2 a_4 a_3 a_6). \end{aligned}$$

This is a cycle of length 5, and by the previous case N must therefore contain a 3-cycle and hence be the whole of A_n .

Case 4 $\omega = \tau(a_1 a_2 a_3)$ where τ is a product of disjoint transpositions. We see that

$$\omega^2 = \tau(a_1 a_2 a_3)\tau(a_1 a_2 a_3) = \tau^2(a_1 a_2 a_3)^2 = (a_1 a_3 a_2)$$

is a 3-cycle, and hence by Lemma 3.42 the subgroup N must again be the entirety of A_n .

Case 5 $\omega = \tau(a_1 a_2)(a_3 a_4)$ where τ is a product of an even number of disjoint transpositions. The conjugate $(a_1 a_2 a_3)\omega(a_1 a_2 a_3)^{-1}$ is in N , and hence so is the product

$$\begin{aligned} \omega^{-1}(a_1 a_2 a_3)\omega(a_1 a_2 a_3)^{-1} \\ &= \tau^{-1}(a_1 a_2)(a_3 a_4)(a_1 a_2 a_3)\tau(a_1 a_2)(a_3 a_4)(a_3 a_2 a_1) \\ &= (a_1 a_3)(a_2 a_4). \end{aligned}$$

Since $n \geq 5$, we can find $a_5 \neq a_1, a_2, a_3, a_4$ and define $\alpha = (a_1 a_3 a_5)$. Then

$$\alpha^{-1}(a_1 a_3)(a_2 a_4)\alpha(a_1 a_3)(a_2 a_4)$$

is an element of N , and

$$\begin{aligned} \alpha^{-1}(a_1 a_3)(a_2 a_4)\alpha(a_1 a_3)(a_2 a_4) \\ &= (a_5 a_3 a_1)(a_1 a_3)(a_2 a_4)(a_1 a_3 a_5)(a_1 a_3)(a_2 a_4) \\ &= (a_1 a_3 a_5) \end{aligned}$$

is a 3-cycle. Hence by Lemma 3.42, $N = A_n$ again.

These are the only possible forms for elements of N , and in each case we deduce that $N = A_n$, and therefore A_n is simple. $\square$

So far, we've discovered two different classes of finite simple groups: the cyclic groups $\mathbb{Z}_p$ of prime order and the alternating groups A_n for $n \geq 5$. A reasonable question to ask at this point is whether there are any more, and it so happens that there are. In one of the longest-running and heroic endeavours in the history of group theory, the finite simple groups have been completely classified:

Theorem 3.44 *Any finite simple group is one of the following four types:*

- (i) Cyclic groups $\mathbb{Z}_p$ of prime order
- (ii) Alternating groups A_n for $n > 4$
- (iii) Finite groups of Lie type
- (iv) Sporadic groups

Proof (Omitted.) $\square$

The proof of this classification theorem took about a hundred mathematicians approximately five decades to complete, between 1955 and 2004,³ and the details are spread through somewhere in excess of ten

³ The proof was declared complete in 1983, but a subtle gap (concerning “quasi-simple groups”) was discovered, which took a further thousand or so pages to patch.

⁴It has been suggested that no single living person fully understands all the details of the proof: the American group theorist Daniel Gorenstein (1923–1992), one of the main driving forces behind the classification, is generally regarded as having had the best overall perspective on the project.

Group	Order
$A_5 \cong A_1(4) = PSL_2(4)$ $\cong A_1(5) = PSL_2(5)$	60
$A_1(7) = PSL_2(7)$ $\cong A_2(2) = PSL_3(2)$	168
$A_6 \cong A_1(9) = PSL_2(9)$	360
$A_1(8) = PSL_2(8)$	504
$A_1(11) = PSL_2(11)$	660
$A_1(13) = PSL_2(13)$	1092
$A_1(17) = PSL_2(17)$	2448
A_7	2520
$A_1(19) = PSL_2(19)$	3420
$A_1(16) = PSL_2(16)$	4080
$A_2(3) = PSL_3(3)$	5616
${}^2A_2(3^3) = PSU_3(3)$	6048
$A_1(23) = PSL_2(23)$	6072
$A_1(25) = PSL_2(25)$	7800
M_{11}	7920
$A_1(27) = PSL_2(27)$	9828

Table 3.11: Nonabelian simple groups of order less than 10 000

thousand pages of journal articles and research monographs.⁴

A full discussion of the third and fourth classes is beyond the scope of this book, but we can look at some examples of finite groups of Lie type now. Table 3.11 lists the nonabelian simple groups with less than 10 000 elements.

Finite groups of Lie type are related to the matrix groups we met in Section 1.2, but we replace $\mathbb{R}$ and $\mathbb{C}$ with finite analogues. We'll study these objects, **finite fields**, in more detail in Chapters 8–11, but the following (incomplete) definition will do for the moment.

Definition 3.45 Let p be prime, and let $\mathbb{F}_p = \{0, \dots, p-1\}$ be the set consisting of the first p non-negative integers, equipped with two binary operations $+$ and $\cdot$ representing, respectively, addition and multiplication modulo p . Then $\mathbb{F}_p$ is the **finite field** of order p .

More generally, given any prime power $q = p^n$ for some $n \in \mathbb{N}$, we can define the finite field $\mathbb{F}_q$, but the general construction is more complicated, and we'll postpone it until Section 11.3.

Here we run into a slight difficulty with notation. The finite groups of Lie type classify neatly into a number of infinite families, many having more than one name. One naming convention, used for some of these groups, intentionally echoes the names used for the corresponding infinite matrix groups. Another widely-used convention (used also for the Coxeter groups we'll meet in Section 5.C) uses the capital letters A – G with numeric subscripts and superscripts; if we're not careful, this could cause confusion with the alternating groups A_n and the dihedral groups D_n . To resolve this, we'll use a slightly different (sans-serif) font in this context so that A_n is the alternating group on n objects, $A_n(q)$ is the finite simple group described next, and A_n is the finite Coxeter group discussed later (and which, confusingly, is isomorphic to the symmetric group S_{n+1}).

In the following examples, unless otherwise stated, q will be a positive power of a prime integer, to ensure that $\mathbb{F}_q$ is indeed a finite field.

Example 3.46 For $n > 1$, let $A_n(q)$ or $PSL_{n+1}(q)$ denote the **projective special linear group** over $\mathbb{F}_q$. That is, the special linear group $SL_{n+1}(\mathbb{F}_q)$ of $(n+1) \times (n+1)$ square matrices, with entries from the order- q finite field $\mathbb{F}_q$ and determinant 1 (calculated in $\mathbb{F}_q$), factored by its centre as described in Example 3.38.

In general,

$$|A_n(q)| = |PSL_{n+1}(q)| = \frac{1}{\gcd(n+1, q-1)} q^{n(n+1)/2} \prod_{i=1}^n (q^{i+1} - 1).$$

So, for example,

$$|A_1(4)| = |PSL_2(4)| = \frac{1}{\gcd(2,3)} 4^{1 \times 2/2} \prod_{i=1}^1 (4^{i+1} - 1) = 4 \times 15 = 60.$$

All of these groups are simple, except for $A_1(2) = PSL_2(2)$ which is isomorphic to the symmetric group S_3 , and $A_1(3) = PSL_2(3)$ which is isomorphic to the alternating group A_4 .

In addition, $A_1(4) = PSL_2(4)$ and $A_1(5) = PSL_2(5)$ are both isomorphic to each other and to the alternating group A_5 ; while $A_1(7) = PSL_2(7)$ and $A_2(2) = PSL_3(2)$ are isomorphic to each other. Also, the group $A_1(9) = PSL_2(9)$ is isomorphic to the alternating group A_6 ; and the group $A_3(2) = PSL_4(2)$ is isomorphic to the alternating group A_8 .

In a similar way, the projective symplectic groups $PSp_n(\mathbb{R})$ yield another family of simple groups:

Example 3.47 For $n > 1$, let $C_n(q)$ or $PSp_{2n}(q)$ denote the **projective symplectic group** over $\mathbb{F}_q$. By analogy with the previous example, we take the group $Sp_{2n}(\mathbb{F}_q)$ of symplectic $2n \times 2n$ matrices with entries from $\mathbb{F}_q$ and factor out by the centre as described in Example 3.38.

We ignore the case $n = 1$ since $Sp_2(\mathbb{F}_q) = SL_2(\mathbb{F}_q)$ and so this is covered by the previous example. All of the groups $C_n(q) = PSp_{2n}(q)$ are simple except for $C_2(2) = PSp_4(2)$ which is isomorphic to the symmetric group S_6 .

In general,

$$|C_n(q)| = |PSp_{2n}(q)| = \frac{1}{\gcd(2, q-1)} q^{n^2} \prod_{i=1}^n (q^{2i} - 1).$$

So, for example,

$$|C_2(2)| = |PSp_4(2)| = \frac{2^4}{\gcd(2,1)} \prod_{i=1}^2 (2^{2i} - 1) = 16 \times 3 \times 15 = 720 = |S_6|.$$

There are three other closely-related families of finite simple groups obtained from the special orthogonal groups $SO_n(\mathbb{F}_q)$, however the details are somewhat involved: in addition to factoring out by their centre to obtain projective versions, we have to consider separately the cases where n is even or odd, and restrict our attention to a subgroup defined in terms of something called the **spinor norm**. This process yields the groups $B_n(q) = O_{2n+1}(q)$ (for $n > 1$), $D_n(q) = O_{2n}^+(q)$ (for $n > 3$) and ${}^2D_n(q^2) = O_{2n}^-(q)$ (for $n > 3$).

More straightforward is the construction of a family of finite simple groups from the projective special unitary groups PSU_n ; this process leads to the groups ${}^2A_n(q^2) = PSU_{n+1}(q)$.

We won't discuss these groups any further, but a complete list appears

Group	Order
$A_n(q) = PSL_{n+1}(q)$	$\frac{1}{\gcd(n+1, q-1)} q^{n(n+1)/2} \prod_{i=1}^n (q^{i+1} - 1)$
$B_n(q) = O_{2n+1}(q)$	$n > 1 \quad \frac{1}{\gcd(2, q-1)} q^{n^2} \prod_{i=1}^n (q^{2i} - 1)$
$C_n(q) = PSp_{2n}(q)$	$n > 2 \quad \frac{1}{\gcd(2, q-1)} q^{n^2} \prod_{i=1}^n (q^{2i} - 1)$
$D_n(q) = O_{2n}^+(q)$	$n > 3 \quad \frac{1}{\gcd(4, q^n-1)} q^{n(n-1)} (q^n - 1) \prod_{i=1}^{n-1} (q^{2i} - 1)$
$E_6(q)$	$\frac{1}{\gcd(3, q-1)} q^{36} (q^{12} - 1)(q^9 - 1)(q^8 - 1)(q^6 - 1)(q^5 - 1)(q^2 - 1)$
$E_7(q)$	$\frac{1}{\gcd(2, q-1)} q^{63} (q^{18} - 1)(q^{14} - 1)(q^{12} - 1)(q^{10} - 1)(q^8 - 1)(q^6 - 1)(q^2 - 1)$
$E_8(q)$	$q^{120} (q^{30} - 1)(q^{24} - 1)(q^{20} - 1)(q^{18} - 1)(q^{14} - 1)(q^{12} - 1)(q^8 - 1)(q^2 - 1)$
$F_4(q)$	$q^{24} (q^{12} - 1)(q^8 - 1)(q^6 - 1)(q^2 - 1)$
$G_2(q)$	$q^6 (q^6 - 1)(q^2 - 1)$
${}^2A_n(q^2) = PSU_{n+1}(q)$	$n > 1 \quad \frac{1}{\gcd(n+1, q+1)} q^{n(n+1)/2} \prod_{i=1}^n (q^{i+1} - (-1)^{i+1})$
${}^2D_n(q^2) = O_{2n}^-(q)$	$n > 3 \quad \frac{1}{\gcd(4, q^n+1)} q^{n(n-1)} (q^n + 1) \prod_{i=1}^{n-1} (q^{2i} - 1)$
${}^2E_6(q^2)$	$\frac{1}{\gcd(3, q+1)} q^{36} (q^{12} - 1)(q^9 + 1)(q^8 - 1)(q^6 - 1)(q^5 + 1)(q^2 - 1)$
${}^3D_4(q^3)$	$q^{12} (q^8 + q^4 + 1)(q^6 - 1)(q^2 - 1)$
${}^2B_2(2^{2n+1})$	$q^2 (q^2 + 1)(q - 1) \quad (\text{where } q = 2^{2n+1})$
${}^2F_4(2^{2n+1})$	$q^{12} (q^6 + 1)(q^4 - 1)(q^3 + 1)(q - 1) \quad (\text{where } q = 2^{2n+1})$
${}^2F_4(2)'$	$17971200 = 2^{11} \times 3^3 \times 5^2 \times 13$
${}^2G_2(3^{2n+1})$	$q^3 (q^3 + 1)(q - 1) \quad (\text{where } q = 3^{2n+1})$

Table 3.12: Simple groups of Lie type

in Table 3.12, all of which are simple except for the following:

$$\begin{aligned}
&A_1(2) = PSL_2(2), \quad A_1(3) = PSL_2(3), \quad B_2(2) = O_5(2), \\
&G_2(2), \quad {}^2A_2(2^2) = PSU_3(2), \quad {}^2B_2(2), \\
&{}^2F_4(2), \quad {}^2G_2(3).
\end{aligned}$$

The **Chevalley groups** $A_n(q)$, $B_n(q)$, $C_n(q)$, $D_n(q)$, $E_6(q)$, $E_7(q)$, $F_4(q)$ and $G_2(q)$ are named after the French mathematician Claude Chevalley (1909–1984); the groups ${}^2A_n(q^2)$, ${}^2D_n(q^2)$, ${}^2E_6(q^2)$ and ${}^3D_4(q^3)$ are called **Steinberg groups** after the Moldovan/American mathematician Robert Steinberg; the groups ${}^2B_2(2^{2n+1})$ are called **Suzuki groups** after the Japanese mathematician Michio Suzuki (1926–1998); the groups ${}^2F_4(2^{2n+1})$ and ${}^2G_2(3^{2n+1})$ are called **Ree groups** after the

Group	Order
M_{11}	7 920
M_{12}	95 040
M_{22}	443 520
M_{23}	10 200 960
M_{24}	244 823 040
J_1	175 560
J_2	604 800
J_3	50 232 960
J_4	86 775 571 046 077 562 880
Co_1	4 157 776 806 543 360 000
Co_2	42 305 421 312 000
Co_3	495 766 656 000
Fi_{22}	64 561 751 654 400
Fi_{23}	4 089 470 473 293 004 800
Fi'_{24}	1 255 205 709 190 661 721 292 800
McL	898 128 000
HS	44 352 000
Suz	448 345 497 600
He	4 030 387 200
Ru	145 926 144 000
HN	273 030 912 000 000
$O'N$	460 815 505 920
Ly	51 765 179 004 000 000
Th	90 745 943 887 872 000
B	4 154 781 481 226 426 191 177 580 544 000 000
M	808 017 424 794 512 875 886 459 904 961 710 757 005 754 368 000 000 000

Table 3.13: Sporadic simple groups

Korean/Canadian mathematician Rimhak Ree (1922–2005), and the group ${}^2F_4(2)'$ (the commutator subgroup of ${}^2F_4(2)$) is called the **Tits group** after the Belgian/French mathematician Jacques Tits.

The fourth class of finite simple groups comprises twenty-six groups which don't fit into the first three classes. These groups are called **sporadic groups** and all have various intriguing properties. They range in size from the Mathieu group M_{11} of order 7920, up to the aptly-named Monster group M with just over 8×10^{53} elements. The sporadic groups and their orders are listed in Table 3.13.

The five **Mathieu groups** M_{11} , M_{12} , M_{22} , M_{23} and M_{24} were discovered in the 1860s and early 1870s by the French mathematician Émile Mathieu (1835–1890), who first studied them as part of his research on multiply transitive permutation groups. All five can be constructed in terms of symmetries of combinatorial objects known as Steiner

systems. The groups M_{23} and M_{24} also happen to be the symmetry groups of the 23- and 24-bit binary error-correcting codes discovered by the Swiss mathematician and information-theorist Marcel Golay (1902–1989); these codes were used by NASA to reliably transmit image data back to earth from the Voyager 1 and 2 space probes during their 1979–1981 flyby missions to Jupiter and Saturn.

The existence of the four **Janko groups** J_1 , J_2 , J_3 and J_4 was predicted by the Croatian mathematician Zvonimir Janko; he constructed J_1 in 1965 and the other three were explicitly constructed (and their existence confirmed) over the next fifteen years by other mathematicians.

The three **Conway groups** Co_1 , Co_2 and Co_3 were discovered by the British mathematician John Conway. They arise as subgroups of the symmetry group Co_0 of a 24-dimensional object called the **Leech lattice** Λ_{24} ; Conway constructed this group during a single twelve-hour period one Saturday in 1964.

The group Co_1 is the quotient of Co_0 by its two-element centre, while the groups Co_2 and Co_3 are the subgroups which leave unchanged certain types of lattice vectors in Λ_{24} . Four other sporadic groups exist as subgroups or quotients of Co_1 : the **McLaughlin group** McL , the **Higman–Sims group** HS , the **Suzuki group** Suz and the second Janko group J_2 (sometimes called the **Hall–Janko group**).

The three **Fischer groups** Fi_{22} , Fi_{23} and Fi'_{24} all arise as groups of 3-**transpositions**, particular types of order-2 permutations where the product of any two has order at most 3. They were discovered in 1971 by the German mathematician Bernd Fischer. The group Fi_{24} is not itself simple, but its commutator subgroup Fi'_{24} is.

The remaining eleven sporadic groups are the **Held group** He , the **Rudvalis group** Ru , the **O’Nan group** $O’N$, the **Harada–Norton group** HN , the **Lyons group** Ly , the **Thompson group** Th , the **Baby Monster group** B and the **Monster group** M .

The last of these contains nineteen of the other twenty-five sporadic groups as subgroups; these nineteen are sometimes called the **happy family**, while the remaining six (J_1 , J_3 , J_4 , Ly , $O’N$ and Ru) are sometimes referred to as **pariahs**.

Summary

In the previous chapter, during our discussion of subgroups, we met the concept of **cosets**.^{5,6} These arose initially from the proof of Cayley's Theorem⁷ and led to Lagrange's Theorem.⁸ It's interesting to ask whether a given left coset gH necessarily coincides with the corresponding right coset Hg . Certainly this will be the case if the larger group G is abelian, but need not be otherwise. We did, however, find a number of cases of nonabelian groups which contained subgroups with this property: the rotation subgroup R_3 in D_3 for example, or the alternating group A_4 in S_4 .⁹

A full study of this situation led us to a relation called **conjugacy**: an element $g \in G$ is **conjugate** to an element $h \in G$ if there exists some element $k \in G$ such that $k*g*k^{-1} = h$.¹⁰ This criterion arises quite naturally in linear algebra: two $n \times n$ matrices are conjugate (or **similar**) exactly when they represent the same linear map $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ relative to a different choice of basis.¹¹

More generally, we find that conjugacy is an equivalence relation,¹² and hence partitions a given group G into a number of equivalence classes, called **conjugacy classes**. The conjugacy classes of D_3 , for example, are $\{e\}$, $\{r, r^2\}$ and $\{m_1, m_2, m_3\}$.¹³ The group $GL_n(\mathbb{R})$ has uncountably many conjugacy classes, but each corresponds to a particular type of Jordan block matrix.¹⁴ If G is abelian then its conjugacy classes are singleton sets, each consisting of one element of G .¹⁵

It transpires that conjugacy classes are the key to understanding when the left and right cosets gH and Hg coincide and when they don't. This correspondence happens exactly when the subgroup H is a union of conjugacy classes.¹⁶ So, for example, any subgroup of an abelian group has this form, since the conjugacy classes all consist of single elements. The rotation subgroup $R_3 = \{e, r, r^2\} = \{e\} \cup \{r, r^2\}$ also has this form, but the reflection subgroup $\{e, m_1\}$ doesn't, so the cosets of the former coincide while those of the latter don't. Equivalently, a subgroup $H < G$ has this property if it's closed under conjugation by any element of G .

We call subgroups with this coincidence property **normal subgroups** and denote them $H \triangleleft G$ or $H \trianglelefteq G$ depending on whether H is a proper subgroup of G or might not be.¹⁷ Subgroups of abelian groups are always normal, and any subgroup of index 2 (such as $R_n < D_n$ or $A_n < S_n$) is also normal.¹⁸ The special linear group $SL_n(\mathbb{R})$ is closed under conjugation by elements of $GL_n(\mathbb{R})$, hence $SL_n(\mathbb{R}) \triangleleft GL_n(\mathbb{R})$.¹⁹ Considering commutativity of individual elements yields some important examples of normal subgroups. The **centre** $Z(G)$ consists

⁵ Definition 2.16, page 50.

⁶ Definition 2.17, page 50.

⁷ Theorem 2.13, page 47.

⁸ Theorem 2.30, page 54.

⁹ Example 3.1, page 70.

¹⁰ Definition 3.2, page 70.

¹¹ Definition 3.3, page 71.

¹² Proposition 3.4, page 72.

¹³ Example 3.5, page 72.

¹⁴ Example 3.6, page 73.

¹⁵ Proposition 3.8, page 73.

¹⁶ Proposition 3.10, page 74.

¹⁷ Definition 3.11, page 75.

¹⁸ Proposition 3.12, page 75.

¹⁹ Example 3.15, page 76.

²⁰ Definition 3.16, page 76.

²¹ Proposition 3.17, page 76.

²² Definition 3.18, page 77.

²³ Definition 3.19, page 78.

²⁴ Proposition 3.21, page 79.

²⁵ Proposition 3.20, page 78.

²⁶ Proposition 3.24, page 83.

²⁷ Definition 3.25, page 84.

²⁸ Example 3.28, page 85.

²⁹ Example 3.30, page 87.

³⁰ Example 3.38, page 90.

³¹ Example 3.32, page 88.

³² Definition 3.39, page 91.

³³ Example 3.40, page 91.

³⁴ Proposition 3.43, page 92.

³⁵ Example 3.46, page 94.

³⁶ Example 3.47, page 95.

³⁷ Table 3.12, page 96.

³⁸ Table 3.13, page 97.

of all elements of G which commute with all other elements of G ,²⁰ this is not only a subgroup of G but happens to be normal as well.²¹ Amongst other things, it gives us a measure of how abelian a particular group is: $Z(G) = G$ exactly when G is abelian.

Similarly, we define the **commutator subgroup** or **derived group** $[G, G]$ or G' to be the subgroup generated by all elements (called **commutators**) of the form $[g, h] = g*h*g^{-1}*h^{-1}$.^{22,23} Again, this is a normal subgroup of G ²⁴ and gives an indication of how abelian G is: $[G, G]$ is trivial if and only if G is abelian, and the larger $[G, G]$ is, the less abelian G is. In particular, for $n \geq 3$, the commutator subgroup $[S_n, S_n]$ is the alternating group A_n .²⁵

The reason we care about normal subgroups isn't just because their left and right cosets line up neatly, it's because their cosets interact in a well-defined way and form a group in their own right.²⁶ A key example of this construction is the formation of the finite cyclic group $\mathbb{Z}_n$ from the cosets of $n\mathbb{Z} = \langle n \rangle$ in $\mathbb{Z}$. We say that $\mathbb{Z}_n$ is the **quotient group** $\mathbb{Z}/n\mathbb{Z}$. More generally, the (left or right) cosets of a normal subgroup $H \trianglelefteq G$ form a quotient group G/H in which the product $(gH)(kH)$ is defined to be the coset $(g*k)H$.²⁷ We also saw that $D_3/R_3 \cong \mathbb{Z}_2$ and $\mathbb{Z}_6/\langle 3 \rangle \cong \mathbb{Z}_3$,²⁸ amongst others.

Factoring a group G by its centre $Z(G)$ often yields useful information about the nonabelianness of a group: the smaller the centre, the larger the quotient $G/Z(G)$; if G is abelian then $Z(G) = G$ and hence $G/Z(G)$ is trivial. On the other hand, $D_3/Z(D_3) = D_3$, so the dihedral group D_3 is in some sense about as nonabelian as we might reasonably expect it to be.²⁹ Factoring the general linear group $GL_n(\mathbb{R})$ by its centre yields the **projective general linear group** $PGL_n(\mathbb{R})$; analogous projective versions of $O_n(\mathbb{R})$, $SO_n(\mathbb{R})$, U_n , SU_n and $Sp_{2n}(\mathbb{R})$ exist too.³⁰

The commutator subgroup is another measure of the nonabelianness of a given group: if it's trivial then the group is abelian. Factoring out by this subgroup (which, as with the centre, we can do because it's a normal subgroup) effectively kills all the noncommutativity in the group, leaving an abelian group, which we'll denote G^{ab} . This process is called **abelianisation**.³¹

Certain groups have no normal subgroups apart from themselves and the trivial subgroup. Another way of looking at this is that they have no proper and nontrivial quotients. We call groups of this type **simple**,³² and they fall into four different classes: cyclic groups $\mathbb{Z}_p$ where p is prime,³³ alternating groups A_n for $n > 4$,³⁴ finite groups of Lie type,^{35,36,37} and twenty-six **sporadic** groups.³⁸

References and further reading

An interesting account of the development of the concept of a quotient group, which emerged towards the end of the 19th century, can be found in the following article:

- J Nicholson, *The development and understanding of the concept of quotient group*, *Historia Mathematica* 20.1 (1993) 68–88

Further details on quaternions, octonions and related topics can be found in the following book and article:

- J H Conway and D A Smith, *On Quaternions and Octonions*, A K Peters (2003)
- J C Baez, *The octonions*, *Bulletin of the American Mathematical Society* 39 (2002) 145–205

Mark Ronan's book *Symmetry and the Monster* provides an accessible account of the programme to classify finite simple groups, written for a lay and nonspecialist audience.

- M Ronan, *Symmetry and the Monster*, Oxford University Press (2006)

Technical details on finite simple groups can be found in the following books, amongst others, intended for graduate students and specialist researchers:

- R Carter, *Simple Groups of Lie Type*, Wiley (1989)
- R A Wilson, *The Finite Simple Groups*, Graduate Texts in Mathematics 251, Springer (2009)
- J H Conway, R T Curtis, S P Norton, R A Parker, and R A Wilson, *Atlas of Finite Groups*, Clarendon Press, Oxford (1985)

A very readable account of the fascinating links between error-correcting codes, information theory, higher-dimensional sphere packings and sporadic simple groups, can be found in Thomas Thompson's book:

- T M Thompson, *From Error-Correcting Codes Through Sphere Packings To Simple Groups*, Carus Mathematical Monographs 21, Mathematical Association of America (1983)

Exercises

Show that $[D_{2n+1}, D_{2n+1}] = R_{2n+1} \cong \mathbb{Z}_{2n+1}$, and $[D_{2n}, D_{2n}] \cong \mathbb{Z}_n$.

4 Homomorphisms

SO FAR, we've met many different types of groups: cyclic, dihedral and other symmetry groups, permutation groups, matrix groups, and several types of simple groups, amongst others. We've studied their internal structure in various ways: decomposition of permutations into disjoint cycles, whether two elements are conjugate, which elements commute with each other, what subgroups a group contains, what happens if we factor by a normal subgroup, and so on.

However, there is an obvious thing we haven't really looked at yet, except in one special case. Groups are essentially just sets equipped with some extra structure. For ordinary sets we can cheerfully define all sorts of functions mapping elements of one set to. We can impose additional conditions (such as injectivity, surjectivity and bijectivity) on these functions, study their images, compose them, and so forth.

So, what happens if we define functions between groups? Can we learn anything useful about the structure of a particular group by examining these functions? Will any function do, or must we restrict our attention to ones satisfying some specific conditions?

4.1 Structure-preserving maps

IT SO HAPPENS that we've already met one special sort of function. In Chapter 1 we saw that some groups were structurally identical, apart from some minor renaming of their elements. For example, the cyclic group $\mathbb{Z}_3$, consisting of the numbers $\{0, 1, 2\}$ under addition modulo 3, is structurally the same as the multiplicative group $C_3 = \{1, -\frac{1}{2} + \frac{\sqrt{3}}{2}i, -\frac{1}{2} - \frac{\sqrt{3}}{2}i\}$ of complex cube roots of unity. And the group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ is effectively the same as the Klein 4-group V_4 and the symmetry group of the rectangle.

The device we introduced to formalise this idea of structural similarity was a particular type of bijective function that respected the group

"You have shown me a strange image, and they are strange prisoners."

"Like ourselves," I replied; "and they see only their own shadows, or the shadows of one another, which the fire throws on the opposite wall of the cave?"

"True," he said; "how could they see anything but the shadows if they were never allowed to move their heads?"

"And of the objects which are being carried in like manner they would only see the shadows?"

"Yes," he said.

"And if they were able to converse with one another, would they not suppose that they were naming what was actually before them?"

"Very true."

— Plato (c.427–347 BC),
The Republic VII (c.380 BC)

Structures are the weapons of the mathematician.

— attributed to Nicolas Bourbaki

structure: an **isomorphism**. It's now time to generalise this concept, and consider non-bijective functions that preserve the group structure.

Definition 4.1 Let $G = (G, *)$ and $H = (H, \bullet)$ be two groups. A function $f: G \rightarrow H$ is a **homomorphism** if $f(g_1 * g_2) = f(g_1) \bullet f(g_2)$ for all $g_1, g_2 \in G$.

We could also consider functions that don't satisfy this structural condition, but it turns out that these aren't very interesting from a group-theoretic point of view. What distinguishes group theory from set theory is precisely this extra structure, so it doesn't really make sense to ignore it. Let's look at some examples.

Example 4.2 The function $f: \mathbb{Z} \rightarrow \mathbb{Z}$ given by $f: n \mapsto 2n$ is a homomorphism. To see this, observe that $f(m+n) = 2(m+n) = 2m+2n = f(m)+f(n)$ for all $m, n \in \mathbb{Z}$.

Example 4.3 The function $f: S_3 \rightarrow S_3$ where $f: \sigma \mapsto (12)\sigma(12)$ is a homomorphism, because

$$f(\sigma\tau) = (12)\sigma\tau(12) = (12)\sigma(12)(12)\tau(12) = f(\sigma)f(\tau)$$

for any $\sigma, \tau \in S_3$.

Example 4.4 The function $f: GL_n(\mathbb{R}) \rightarrow GL_n(\mathbb{R})$ where $A \mapsto \frac{1}{\det A} A$ is a homomorphism, since

$$f(AB) = \frac{1}{\det AB} AB = \frac{1}{\det A \det B} AB = \left(\frac{1}{\det A} A\right) \left(\frac{1}{\det B} B\right) = f(A)f(B)$$

for any matrices $A, B \in GL_n(\mathbb{R})$.

In these examples, the homomorphism maps from a group to itself: the domain and codomain are the same. We call this type of homomorphism an **endomorphism**. If, like Example 4.3, it is bijective, an isomorphism from a group to itself, we call it an **automorphism**.

Example 4.5 The function $f: \mathbb{Z} \rightarrow \mathbb{Z}_5$ where $n \mapsto [n]_5$, the remainder of n modulo 5, is a homomorphism:

$$f(m+n) = [m+n]_5 = [m]_5 + [n]_5 = f(m) + f(n)$$

Example 4.6 The function $f: GL_n(\mathbb{R}) \rightarrow \mathbb{R}_*^\times$ which maps a nonsingular, real, $n \times n$ matrix A to its determinant $\det A$, is a homomorphism, since $\det(AB) = \det A \det B$ for any matrices $A, B \in GL_n(\mathbb{R})$.

Parity of permutations gives a homomorphism into $\mathbb{Z}_2$:

Example 4.7 Let $f: S_n \rightarrow \mathbb{Z}_2$, where

$$f(\sigma) = \begin{cases} 0 & \text{if } \sigma \text{ is even,} \\ 1 & \text{if } \sigma \text{ is odd.} \end{cases}$$

Then f is a homomorphism, by the proof of Proposition 1.61.

In these three examples, the homomorphism f is surjective; we call a surjective homomorphism an **epimorphism**. Sometimes we'll denote this with a double-headed arrow: $f: G \twoheadrightarrow H$.

Example 4.8 Let $f: 3\mathbb{Z} \rightarrow \mathbb{Z}$ be the obvious “divide by 3” function $n \mapsto \frac{n}{3}$. This is a homomorphism, since $f(m+n) = \frac{m+n}{3} = \frac{m}{3} + \frac{n}{3} = f(m) + f(n)$.

Example 4.9 Let $f: A_n \rightarrow S_n$ be the obvious inclusion function. This is a homomorphism, since A_n is a subgroup of S_n .

Example 4.10 Let $f: \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2$ by $n \mapsto (n, n)$ for either $n \in \mathbb{Z}_2$. This is a homomorphism, because $f(m+n) = (m+n, m+n) = (m, m) + (n, n) = f(m) + f(n)$ for any $m, n \in \mathbb{Z}_2$.

These three examples are of injective homomorphisms, or **monomorphisms**. Sometimes we'll use an arrow with a tail to denote an injective homomorphism: $f: G \rightarrowtail H$. In fact, Example 4.8 happens to be not just injective but surjective as well, and hence an isomorphism.

Given two functions $f: A \rightarrow B$ and $g: B \rightarrow C$, we can chain them together by plugging the output of f into the input of g to form the **composite** function $g \circ f: A \rightarrow C$ where $(g \circ f)(a) = g(f(a))$ for all $a \in A$. Remember that it's important to get the order right: by convention we compose functions on the left. In fact, in general, it doesn't even make sense to discuss $f \circ g$ unless the image of g lies within the domain of f . (Compare this with the strongly analogous case of matrix multiplication: even if the product AB is defined, it might not be the case that BA is defined as well.)

The next proposition is a bit of technical book-keeping to check that composing two homomorphisms gives another homomorphism rather than just an ordinary function.

Proposition 4.11 Let $G = (G, *)$, $H = (H, \bullet)$ and $K = (K, \star)$ be groups, and let $\phi: G \rightarrow H$ and $\psi: H \rightarrow K$ be homomorphisms. Then the composite function $(\psi \circ \phi): G \rightarrow K$ is a homomorphism.

Proof For any $g_1, g_2 \in G$, we have

$$\begin{aligned} (\psi \circ \phi)(g_1 * g_2) &= \psi(\phi(g_1 * g_2)) \\ &= \psi(\phi(g_1) \bullet \phi(g_2)) && (\phi \text{ is a homomorphism}) \\ &= \psi(\phi(g_1)) \star \psi(\phi(g_2)) && (\psi \text{ is a homomorphism}) \\ &= (\psi \circ \phi)(g_1) \star (\psi \circ \phi)(g_2). \end{aligned}$$

Hence $\psi \circ \phi$ is a homomorphism. □

There are some special homomorphisms worth mentioning at this point, which will recur in various contexts later in the book.

Example 4.12 The **identity** homomorphism $\text{id}_G: G \rightarrow G$ for a group G is, as its name suggests, the homomorphism which maps every element $g \in G$ to itself. (Often we will omit the subscript G if the context is clear.) In particular, the identity homomorphism has the special property that it has no effect on any other homomorphism it's composed with. More precisely, given a homomorphism $f: G \rightarrow H$, we have

$$f \circ \text{id}_G = f = \text{id}_H \circ f,$$

and indeed it's possible to define the identity homomorphism as the unique homomorphism which has this property for a given group. The identity homomorphism is bijective, and hence an isomorphism.

Example 4.13 Another special homomorphism we'll meet from time to time is the **zero** homomorphism $z: G \rightarrow H$ for any groups G and H . This is the unique homomorphism which maps every element in G to the identity element $e \in H$.

Example 4.9 illustrates another important case, which the next example generalises.

Example 4.14 For any subgroup $H < G$ there is a unique **inclusion** homomorphism $i: H \rightarrow G$ which maps every element of H to itself in G . (Sometimes we use a hooked arrow for inclusions instead: $i: G \hookrightarrow H$.) This homomorphism is necessarily injective.

Direct products of groups yield two more types of homomorphisms:

Example 4.15 A direct product $G = G_1 \times \cdots \times G_n$ of groups has a collection of **projection** homomorphisms $p_1: G \rightarrow G_1, \dots, p_n: G \rightarrow G_n$ which map everything onto the corresponding factor. That is, we define $p_i: G \rightarrow G_i$ by $p_i(g_1, \dots, g_n) = g_i \in G_i$. These homomorphisms are necessarily surjective.

For example:

$p_1: \mathbb{Z}_2 \oplus \mathbb{Z}_3 \rightarrow \mathbb{Z}_2$	$p_2: \mathbb{Z}_2 \oplus \mathbb{Z}_3 \rightarrow \mathbb{Z}_3$
$(0, 0), (0, 1), (0, 2) \mapsto 0$	$(0, 0), (1, 0) \mapsto 0$
$(1, 0), (1, 1), (1, 2) \mapsto 1$	$(0, 1), (1, 1) \mapsto 1$
	$(0, 2), (1, 2) \mapsto 2$

Example 4.16 We can also define a sequence of inclusion homomorphisms $i_1: G_1 \rightarrow G, \dots, i_n: G_n \rightarrow G$ where $g \mapsto (e, \dots, e, g, e, \dots, e)$ for all $g \in G_j$. That is, we map an element $g \in G_j$ to the element of the direct product $G = G_1 \times \cdots \times G_n$ with g in the j th place, and e everywhere else. These homomorphisms are injective.

For example:

$$\begin{array}{ll}
 i_1: \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_3 & i_2: \mathbb{Z}_3 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_3 \\
 0 \mapsto (0, 0) & 0 \mapsto (0, 0) \\
 1 \mapsto (1, 0) & 1 \mapsto (0, 1) \\
 & 2 \mapsto (0, 2)
 \end{array}$$

For isometry groups or other groups with some kind of geometric context, we can often use our geometric intuition to devise interesting homomorphisms. The following example illustrates this approach.

Example 4.17 The dihedral groups D_3 and D_6 are the symmetry groups of, respectively, an equilateral triangle and a regular hexagon. Just as we can inscribe an equilateral triangle inside a regular hexagon (see Figure 4.1), we can define a homomorphism which maps D_3 into D_6 :

$$\begin{array}{lll}
 e \mapsto e & r \mapsto r^2 & r^2 \mapsto r^4 \\
 m_1 \mapsto m_2 & m_2 \mapsto m_4 & m_3 \mapsto m_6
 \end{array}$$

More generally, we can define an inclusion homomorphism $D_n \rightarrow D_{2n}$ by mapping

$$e \mapsto e \quad r^k \mapsto r^{2k} \quad m_k \mapsto m_{2k}$$

for all $k \in \{1, \dots, n\}$.

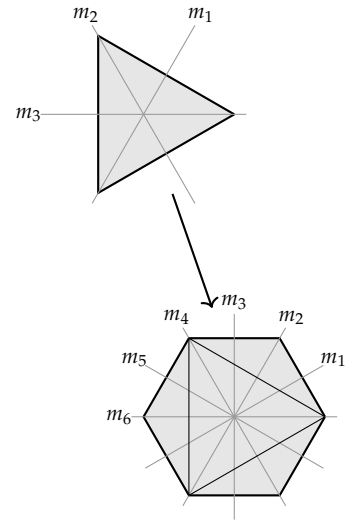


Figure 4.1: Embedding D_3 into D_6

In the last chapter we learned about quotient groups: the formation of a new group G/H from the cosets of a normal subgroup H of G . We found that if H is normal in G , then we can use the group structure of G to define a group structure on the set of cosets of H in G . Another way of looking at this is that the quotienting process in some way respects the group structure of G . The following example uses this insight to define a homomorphism $q: G \rightarrow G/H$.

Example 4.18 Let G be a group, and H be a normal subgroup of G . Define the **quotient homomorphism** $q: G \rightarrow G/H$ to be the function mapping a given element $g \in G$ to the coset $gH \in G/H$. This is a homomorphism since

$$q(g_1 * g_2) = (g_1 * g_2)H = (g_1H)(g_2H) = q(g_1)q(g_2)$$

for any $g_1, g_2 \in G$. It is also surjective, since every coset in G/H is mapped to by at least one element of G .

Some familiar functions can be described as quotient homomorphisms:

Example 4.19 In Example 3.35 we saw $\mathbb{C}_*^\times / \mathbb{R}_+^\times \cong \mathbb{C}$. The quotient homomorphism is the argument map $\arg: \mathbb{C}_*^\times \rightarrow \mathbb{C}$ which assigns to each nonzero complex number z its argument $\arg(z) \in [0, 2\pi)$.

In Example 3.36 we saw that $\mathbb{C}_*^\times / C \cong \mathbb{R}_+^\times$. The quotient homomorphism $\mathbb{C}_*^\times \rightarrow \mathbb{R}_+^\times$ is precisely the modulus map $|-|: \mathbb{C}_*^\times \rightarrow \mathbb{R}_+^\times$ which assigns to each nonzero complex number z its modulus $|z|$.

Example 4.20 At the beginning of Section 3.2 we discussed the construction of the cyclic group $\mathbb{Z}_n$ by factoring $\mathbb{Z}$ by its subgroup $n\mathbb{Z}$. The quotient homomorphism $q: \mathbb{Z} \rightarrow \mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$ in this case is the map $m \mapsto [m]_n$ which maps an integer m to its remainder modulo n .

The following example considers the special case where we factor a group G by its commutator subgroup $G' = [G, G]$:

Example 4.21 Let $\alpha: G \rightarrow G^{\text{ab}} = G/[G, G]$ be the quotient homomorphism from G to its abelianisation. More precisely, $\alpha(g) = g[G, G]$ for any $g \in G$; that is, α maps an element g to the corresponding coset of the commutator subgroup. We call this the **abelianisation homomorphism**; if G is already abelian then α is just the identity map id_G .

The abelianisation homomorphism has another important property: given a homomorphism $f: G \rightarrow H$, abelianisation not only yields groups G^{ab} and H^{ab} , but also a homomorphism $f^{\text{ab}}: G^{\text{ab}} \rightarrow H^{\text{ab}}$.

Proposition 4.22 For any group homomorphism $f: G \rightarrow H$, there is a well-defined homomorphism $f^{\text{ab}}: G^{\text{ab}} \rightarrow H^{\text{ab}}$, called the **induced homomorphism**, given by the mapping $g[G, G] \mapsto f(g)[H, H]$, which interacts neatly with the abelianisation maps $\alpha_G: G \rightarrow G^{\text{ab}}$ and $\alpha_H: H \rightarrow H^{\text{ab}}$ in the sense that $\alpha_H \circ f = f^{\text{ab}} \circ \alpha_G$.

Proof We need to check that f^{ab} is well-defined as a function from G^{ab} to H^{ab} , and that it satisfies the homomorphism condition.

What we mean by the first of these is that it's not enough to know that for any $g \in G$, the coset $g[G, G]$ in G^{ab} is mapped to $f(g)[H, H]$, we need to check that for any other $k \in g[G, G]$, the coset $k[G, G] = g[G, G]$ also gets mapped to $f(g)[H, H]$.

Suppose, then, that $k \in g[G, G]$. This implies that there exists some $h \in [G, G]$ such that $k = g * h$. Then

$$\begin{aligned} f(k)[H, H] &= f(g * h)[H, H] \\ &= (f(g) \bullet f(h))[H, H] && (f \text{ is a homomorphism}) \\ &= (f(g)[H, H])(f(h)[H, H]) && (\alpha_H \text{ is a homomorphism}) \\ &= f(g)[H, H] && (f(h)[H, H] = [H, H]) \end{aligned}$$

and so f^{ab} is well-defined.

That f^{ab} is a homomorphism follows from the fact that f and α_H are, and hence so is their composition. $\alpha_H \circ f$. $\square$

The homomorphism f^{ab} is called the **induced homomorphism**, and is best illustrated by means of another example.

Example 4.23 We saw in Example 3.32 that $D_3^{ab} \cong \mathbb{Z}_2$, and it so happens that $D_6^{ab} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Example 4.17 gave us an inclusion homomorphism $i: D_3 \rightarrow D_6$, so putting all these together should yield an induced homomorphism $i^{ab}: \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

We can depict this situation with a commutative diagram:

$$\begin{array}{ccc} D_3 & \xrightarrow{i} & D_6 \\ \alpha_{D_3} \downarrow & & \downarrow \alpha_{D_6} \\ \mathbb{Z}_2 & \xrightarrow{i^{ab}} & \mathbb{Z}_2 \oplus \mathbb{Z}_2 \end{array}$$

To properly determine the induced homomorphism i^{ab} , we follow the various elements of D_3 around the commutative square. In other words, we choose an element g of D_3 in the top left-hand corner and then calculate its image $\alpha_{D_3}(g)$ with the left-hand vertical map. Then we look at what happens to the same element g if we follow it along the top edge and down the right-hand side of the square, to obtain the composite $\alpha_{D_6}(i(g))$. We define the induced homomorphism i^{ab} to be the unique homomorphism that maps $\alpha_{D_3}(g) \mapsto \alpha_{D_6}(i(g))$; this is guaranteed to be well-defined by the above proposition.

$$\begin{array}{ccc} e \mapsto e & r^k \mapsto r^{2k} & m_k \mapsto m_{2k} \\ \downarrow & \downarrow & \downarrow \\ 0 \mapsto (0,0) & 0 \mapsto (0,0) & 1 \mapsto (1,0) \end{array}$$

From this, we can see that the induced homomorphism i^{ab} is the inclusion homomorphism $i_1: \mathbb{Z}_2 \rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2$ mapping $D_3^{ab} = \mathbb{Z}_2$ into the first coordinate of $D_6^{ab} = \mathbb{Z}_2 \oplus \mathbb{Z}_2$:

$$\begin{aligned} i^{ab}: \mathbb{Z}_2 &\rightarrow \mathbb{Z}_2 \oplus \mathbb{Z}_2 \\ 0 &\mapsto (0,0) \\ 1 &\mapsto (1,0) \end{aligned}$$

We can represent the induced homomorphism f^{ab} by means of the following square of groups and homomorphisms:

$$\begin{array}{ccc} G & \xrightarrow{f} & H \\ \alpha_G \downarrow & & \downarrow \alpha_H \\ G^{ab} & \xrightarrow{f^{ab}} & H^{ab} \end{array}$$

There are two ways of getting from G to H^{ab} : do f then α_H , or do α_G then f^{ab} . The point is that f^{ab} is the unique homomorphism $G^{ab} \rightarrow H^{ab}$ that ensures these two paths are equivalent, in the sense that $(\alpha_H \circ f)(g) = (f^{ab} \circ \alpha_G)(g)$ for all $g \in G$.

We call a diagram of this type a **commutative diagram**.

Notice that for all of the homomorphisms $f: G \rightarrow H$ we've seen so far, the identity e_G of G is mapped to the identity e_H in H . There may be other elements in G which are also mapped to e_H depending on the exact definition of f , and we'll look further into that in the next section, but so far all of the examples we've looked at have been such that $f(e_G) = e_H$. The following proposition confirms that this isn't a coincidence, and that all homomorphisms have this property. In addition, for all of the examples so far, we find that for any element $g \in G$, the inverse g^{-1} of g is mapped to the inverse of whatever

g is mapped to; more formally, $f(g^{-1}) = f(g)^{-1}$. This is also not a coincidence: the point of a homomorphism is that it's a function which in some way preserves the structure of the group in question, so it makes sense that the inverse of an element would be mapped to the inverse of whatever the original element is mapped to.

Proposition 4.24 *Let $f: G \rightarrow H$ be a homomorphism. then $f(e_G) = e_H$, and $f(g^{-1}) = f(g)^{-1}$ for any $g \in G$.*

Proof Suppose that $f(e_G) = h$ for some element $h \in H$. Then

$$h = f(e_G) = f(e_G * e_G) = f(e_G) \bullet f(e_G) = h \bullet h.$$

Hence, by the cancellation laws (Proposition 1.15) $h = e_H$.

Similarly, for any $g \in G$,

$$\begin{aligned} f(g) \bullet f(g^{-1}) &= f(g * g^{-1}) = f(e_G) = e_H \\ \text{and } f(g^{-1}) \bullet f(g) &= f(g^{-1} * g) = f(e_G) = e_H. \end{aligned}$$

What this tells us is that $f(g^{-1})$ acts as a left and right inverse for $f(g)$, or in other words $f(g^{-1}) = f(g)^{-1}$. $\square$

When you want to see if your picture corresponds throughout with the objects you have drawn from nature, take a mirror and look in that at the reflection of the real things, and compare the reflected image with your picture, and consider whether the subject of the two images duly corresponds in both, particularly studying the mirror.

— Leonardo da Vinci (1452–1519),
The Practice of Painting

4.2 Kernels and images

RECALL THAT FOR ANY FUNCTION $f: X \rightarrow Y$ we can define the **image** $\text{im}(f)$ of f . This is a subset of the codomain Y and is defined to consist of those elements of Y that are mapped to by at least one element of the domain X . That is,

$$\text{im}(f) = \{y \in Y : y = f(x) \text{ for some } x \in X\}.$$

It's natural to ask whether the image of a homomorphism $f: G \rightarrow H$ has any interesting properties. We know that from Proposition 4.24 that $\text{im}(f)$ must contain the identity e_H as well as a full complement of inverses; that is, for any $h \in \text{im}(f)$ we know that $h^{-1} \in \text{im}(f)$ too. We now find ourselves moving inexorably towards the following proposition.

Proposition 4.25 *Let $f: G \rightarrow H$ be a group homomorphism. Then the image $\text{im}(f)$ is a subgroup of H .*

Proof We know from Proposition 4.24 that $e_H \in \text{im}(f)$ and also that $h^{-1} \in \text{im}(f)$ for any $h \in \text{im}(f)$. All we need to do now is show that $\text{im}(f)$ is closed under the group multiplication operation in H .

Suppose, therefore, that $h_1 = f(g_1)$ and $h_2 = f(g_2)$ for some elements $g_1, g_2 \in G$. Then

$$h_1 \bullet h_2 = f(g_1) \bullet f(g_2) = f(g_1 * g_2) \in \text{im}(f).$$

Hence $\text{im}(f)$ is indeed closed under the group multiplication operation in H , and is thus a subgroup of H . $\square$

So $\text{im}(f)$ is a subgroup of H . Is there anything else we can say about it? In the last chapter we spent some time and effort studying normal subgroups and using them to construct quotient groups. Is the image of a homomorphism a normal subgroup of its codomain? It certainly will be if the codomain is an abelian group (since all subgroups of abelian groups are normal) but as the following example shows, $\text{im}(f)$ need not be a normal subgroup if the codomain of f is nonabelian.

Example 4.26 Let $f: D_3 \rightarrow D_3$ such that

$$\begin{aligned} e, r, r^2 &\longmapsto e, \\ m_1, m_2, m_3 &\longmapsto m_1. \end{aligned}$$

This is a homomorphism (check this) but its image $\text{im}(f) = \{e, m_1\}$ isn't a normal subgroup of D_3 .

The following example considers the image of a particular injective homomorphism we met in the previous section.

Example 4.27 In Example 4.17 we constructed an injective homomorphism $f: D_3 \rightarrow D_6$. The image $\text{im}(f)$ is a six-element subgroup of D_6 consisting of the elements $\{e, r^2, r^4, m_2, m_4, m_6\}$ which, on further inspection, turns out to be isomorphic to D_3 .

This is true in general: given an injective homomorphism $f: G \rightarrow H$, the image $\text{im}(f)$ is isomorphic to the domain G . An explicit isomorphism can be constructed by restricting the codomain of f to just the image, thereby making the already injective homomorphism f surjective as well.

Having looked at what happens at one end (the codomain) of a homomorphism, we now turn our attention to the other end (the domain) via a short detour into linear algebra.

An important application of linear algebra is the solution of systems of simultaneous linear equations. For example, we can encode the system

$$\begin{aligned} x - 3y + 4z - 2w &= 5 \\ 2y + 5z + w &= 2 \\ y - 3z &= 4 \end{aligned} \tag{4.1}$$

as the matrix equation

$$\begin{bmatrix} 1 & -3 & 4 & -2 \\ 0 & 2 & 5 & 1 \\ 0 & 1 & -3 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \\ 4 \end{bmatrix} \tag{4.2}$$

We can then restate the original problem (4.1) as an attempt to charac-

¹ We are tacitly assuming that (4.2) actually has a solution; in this case it does. Homogeneous systems are always solvable; the original inhomogeneous system (4.2) is solvable if the vector

$$\begin{bmatrix} 5 \\ 2 \\ 4 \end{bmatrix}$$

belongs to the image of A , which it does.

terise the vectors in $\mathbb{R}^4$ which get mapped to the vector

$$\begin{bmatrix} 5 \\ 2 \\ 4 \end{bmatrix}$$

in $\mathbb{R}^3$ by the linear map represented by the matrix

$$A = \begin{bmatrix} 1 & -3 & 4 & -2 \\ 0 & 2 & 5 & 1 \\ 0 & 1 & -3 & 0 \end{bmatrix}.$$

Suppose that $\mathbf{v}$ is a solution of (4.2).¹ Then any other solution of (4.2) can be expressed as a sum $\mathbf{u} + \mathbf{v}$ where $\mathbf{u}$ is a solution of the associated **homogeneous system**

$$\begin{bmatrix} 1 & -3 & 4 & -2 \\ 0 & 2 & 5 & 1 \\ 0 & 1 & -3 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (4.3)$$

The solutions of the homogeneous system form a **subspace** of the vector space $\mathbb{R}^4$ which we call the **null space** or **kernel** of (the linear map represented by) the matrix A . In this case, the general solution of (4.2) takes the form

$$\begin{bmatrix} 5 \\ 4 \\ 0 \\ -6 \end{bmatrix} + t \begin{bmatrix} -17 \\ 3 \\ 1 \\ -11 \end{bmatrix}$$

where $t \in \mathbb{R}$ is a real parameter. Here we can see that the kernel of A is a one-dimensional subspace of $\mathbb{R}^4$.

So, in linear algebra we can derive some useful information about a linear map (and thereby, perhaps, solve the actual problem under investigation) by characterising which elements of the domain get mapped to the zero vector in the codomain; that is, by studying the kernel of the map.

We now define the analogous concept for group homomorphisms:

Definition 4.28 Let $f: G \rightarrow H$ be a group homomorphism. Then the **kernel** $\ker(f)$ of f is defined to be

$$\ker(f) = \{g \in G : f(g) = e_H\}.$$

That is, the elements of G that are mapped by f to the identity e_H in H .

Looking again at the homomorphism $f: D_3 \rightarrow D_3$ from Example 4.26, we can see that the kernel of f is the set $\{e, r, r^2\}$, which happens to be the rotation subgroup R_3 of D_3 .

Just as the image of a homomorphism is a subgroup of the codomain, and the kernel and image of a linear map are vector subspaces of, respectively, the domain and codomain, it seems reasonable to ask whether the kernel of a group homomorphism is necessarily a subgroup of the domain. But the kernel of the homomorphism in Example 4.26 isn't just a subgroup of the domain, it's a normal subgroup. Might it be that the kernel is always a normal subgroup of the domain? The answer to this question is yes, as the following proposition shows.

Proposition 4.29 *Suppose $f: G \rightarrow H$ is a group homomorphism. Then $\ker(f)$ is a normal subgroup of G . Furthermore, any normal subgroup of a group G can be expressed as the kernel of some suitable homomorphism*

Proof Clearly the identity $e_G \in \ker(f)$, since $f(e_G) = e_H$ by Proposition 4.24. Given any $g \in \ker(f)$ we see that

$$f(g^{-1}) = f(g)^{-1} = e_H^{-1} = e_H$$

and so $g^{-1} \in \ker(f)$ as well. Given any $g_1, g_2 \in \ker(f)$ we have

$$f(g_1 * g_2) = f(g_1) \bullet f(g_2) = e_H \bullet e_H = e_H$$

and so $g_1 * g_2 \in \ker(f)$ as well. Hence $\ker(f)$ is a subgroup of G : it contains the identity e_G , it contains all necessary inverses, and it's closed under the group multiplication operation $*$ in G .

For $\ker(f)$ to be a normal subgroup of G we need it to be closed under conjugation by any element of G . So let $k \in \ker(f)$ be some element of the kernel, and let g be some arbitrary element of G . Then

$$\begin{aligned} f(g * k * g^{-1}) &= f(g) * f(k) * f(g^{-1}) \\ &= f(g) \bullet e_H \bullet f(g)^{-1} \\ &= f(g) \bullet f(g)^{-1} \\ &= e_H \end{aligned}$$

and hence the conjugate $g * k * g^{-1} \in \ker(f)$ as required, so $\ker(f)$ is a normal subgroup of G .

To show the second part of the proposition, that any normal subgroup $N \trianglelefteq G$ can be expressed as the kernel of a suitable homomorphism, we make use of the quotient homomorphism introduced in the previous section. Let $q: G \rightarrow G/N$ be this homomorphism. Then N is exactly that part of G which is mapped to the identity in the quotient group G/N , and so $N = \ker(q)$. $\square$

We'll make use of this fact in the next section, but for the moment we'll look at a few examples.

Example 4.30 Let $f: S_n \rightarrow \mathbb{Z}_2$ be the homomorphism which maps even permutations to 0 and odd permutations to 1. Then $\ker(f) = A_n$, the alternating group of even permutations. This is known to be normal in S_n (see Example 3.33).

Example 4.31 Example 4.26 provided a homomorphism $f: D_3 \rightarrow D_3$ whose kernel is the rotation subgroup $R_3 \triangleleft D_3$. By Proposition 4.29 we can now confidently state that there is no homomorphism $f: D_3 \rightarrow H$, where $H = D_3$ or any other group, such that $\ker(f) = \{e, m_1\}$, since this isn't a normal subgroup of D_3 .

In fact, no matter what group H is, there are only three possibilities

for the kernel of a homomorphism $f: D_3 \rightarrow H$. Either $\ker(f) = D_3$, in which case f is the zero homomorphism (see Example 4.13) which maps everything in D_3 to $e_H \in H$, or $\ker(f)$ is the trivial subgroup $\{e_G\}$, in which case f is an injective homomorphism, or $\ker(f) = R_3$.

Example 4.32 By the same argument used in the previous example, if $n > 4$ then there are no homomorphisms $f: A_n \rightarrow H$ for any group H , other than the zero homomorphism which maps all of A_n to $e_H \in H$, or an injective homomorphism mapping A_n isomorphically to some subgroup of H . This follows from the simplicity of A_n for $n > 4$ (see Proposition 3.43).

Example 4.33 We know from Example 4.6 that the determinant map $\det: GL_n(\mathbb{R}) \rightarrow \mathbb{R}_*^\times$ is a homomorphism. The kernel of this homomorphism is precisely the subgroup of $n \times n$ matrices with determinant 1, or in other words the special linear group $SL_n(\mathbb{R})$. We already know that $SL_n(\mathbb{R})$ is a normal subgroup of $GL_n(\mathbb{R})$ (see Example 3.15) but if we didn't, this would confirm it for us.

Example 4.34 Let $f: \mathbb{Z} \rightarrow \mathbb{Z}$ be the homomorphism which maps an integer m to its modulo- n remainder $[m]_n$. Then $\ker(f) = n\mathbb{Z}$.

The injective homomorphism $f: D_3 \rightarrow D_3$ in Example 4.26 has a trivial kernel; that is, $\ker(f) = \{e\}$. In Examples 4.31 and 4.32 we remarked that the homomorphisms with trivial kernel were those which mapped the group in question injectively (in fact, isomorphically) to a subgroup of their codomain. The next proposition confirms that this is a general phenomenon.

Proposition 4.35 *Let $f: G \rightarrow H$ be a group homomorphism. Then f is injective if and only if $\ker(f) = \{e_G\}$.*

Proof If f is injective then $\ker(f)$ clearly has to be trivial. For, if there exist $g_1, g_2 \in G$ with $f(g_1) = f(g_2) = e_H$, then the injectivity of f requires that $g_1 = g_2$, and Proposition 4.24 implies that $g_1 = g_2 = e_G$. Hence $\ker(f) = \{e_G\}$.

The converse is only very slightly more involved. Suppose that $\ker(f) = \{e_G\}$ and that $g_1, g_2 \in G$ are such that $f(g_1) = f(g_2)$. Then

$$e_H = f(g_1) \bullet f(g_2)^{-1} = f(g_1) \bullet f(g_2^{-1}) = f(g_1 * g_2^{-1}).$$

So $g_1 * g_2^{-1}$ is in the kernel of f . But $\ker(f)$ is trivial, so $g_1 * g_2^{-1} = e_G$ and hence $g_1 = g_2$. Therefore f is injective. $\square$

The kernel of a homomorphism $f: G \rightarrow H$ is the subset (actually, by Proposition 4.29, the normal subgroup) of G consisting of all elements which f maps to the identity e_H in H . We can generalise this notion a little bit by considering the elements of G which get mapped to other

elements of H .

Definition 4.36 Let $f: G \rightarrow H$ be a group homomorphism, and let h be some element of H . We define the **inverse image** or **preimage** of h to be the set

$$f^{-1}(h) = \{g \in G : f(g) = h\} \subseteq G$$

of elements of G which f maps to the element h .

If K is a subset of H , then the **inverse image** or **preimage** of K is the set

$$f^{-1}(K) = \{g \in G : f(g) \in K\} \subseteq G$$

of elements of G which f maps to some element of K .

In particular, $\ker(f) = f^{-1}(e_H)$, and for any single element $h \in H$ we have $f^{-1}(h) = f^{-1}(\{h\})$. Also, if f is injective, then $f^{-1}(h)$ will consist of a single element, and if f is not surjective, then there will exist at least one element $h \in H$ for which $f^{-1}(h) = \emptyset$.

What happens if we look at the inverse image, not just of a subset, but of a subgroup of H ? The following proposition answers this question.

Proposition 4.37 Let $f: G \rightarrow H$ be a group homomorphism, and let $K \leq H$ be a subgroup of H . Then the inverse image $f^{-1}(K)$ is a subgroup of G . If $K \trianglelefteq H$ is a normal subgroup of H , then the inverse image $f^{-1}(K)$ is a normal subgroup of G .

Proof The inverse image $f^{-1}(K)$ certainly contains the identity e_G , since $f(e_G) = e_H \in K$. If $g \in f^{-1}(K)$ then $g^{-1} \in f^{-1}(K)$ too, since $f(g^{-1}) = f(g)^{-1} \in K$, because K is a subgroup of H .

Suppose $g_1, g_2 \in f^{-1}(K)$. This means that $f(g_1)$ and $f(g_2)$ are in K , and since K is a subgroup of H , the product $f(g_1) \bullet f(g_2) \in K$ as well. But $f(g_1) \bullet f(g_2) = f(g_1 * g_2)$ since f is a homomorphism, and hence $g_1 * g_2 \in f^{-1}(K)$ as well. Thus $f^{-1}(K)$ is a subgroup of G .

If K is a normal subgroup of H then for any $g_1, g_2 \in f^{-1}(K)$ the conjugate $f(g_1) \bullet f(g_2) \bullet f(g_1)^{-1}$ is also in K , and since

$$f(g_1) \bullet f(g_2) \bullet f(g_1)^{-1} = f(g_1) \bullet f(g_2) \bullet f(g_1^{-1}) = f(g_1 * g_2 * g_1^{-1})$$

we see that the conjugate $g_1 * g_2 * g_1^{-1}$ is also in the inverse image $f^{-1}(K)$, and so $f^{-1}(K)$ is normal in G . $\square$

More generally, for a group homomorphism $f: G \rightarrow H$, the preimage of any single element $h \in H$ is either the empty set $\emptyset$ (if $h \in H \setminus \text{im}(f)$) or a coset of $\ker(f)$ in G .

Proposition 4.38 Let $f: G \rightarrow H$ be a group homomorphism with kernel $K = \ker(f)$, and let $h \in H$ be such that $f(g) = h$ for some $g \in G$. Then $f^{-1}(h) = gK$.

Proof To show $f^{-1}(h) = gK$ we follow the standard approach to proving equality of two sets: we show each is a subset of the other.

Firstly, we show that $f^{-1}(h) \subseteq gK$. Consider some element $x \in f^{-1}(h)$, so $f(x) = h$. Then $f(x) = f(g)$, so

$$e_H = f(g)^{-1} \bullet f(x) = f(g^{-1}) \bullet f(x) = f(g^{-1} * x).$$

This means that $g^{-1} * x \in \ker(f) = K$, and hence by (the left-handed version of) Proposition 2.26 we have $x \in gK$ and thus $f^{-1}(h) \subseteq gK$.

Secondly, to show that $gK \subseteq f^{-1}(h)$ let $x \in gK$. Then $x = g * k$ for some element $k \in K$, and therefore

$$f(x) = f(g * k) = f(g) \bullet f(k) = h \bullet e_H = h.$$

Thus $x \in f^{-1}(h)$, so $gK \subseteq f^{-1}(h)$.

These two inclusions combine to give $f^{-1}(h) = gK$ as required. $\square$

From this we get the following neat little corollary, which will come in useful in a little while.

Corollary 4.39 *Let $f: G \rightarrow H$ be a group homomorphism with kernel $K = \ker(f)$. Then if $|K| = n$, the homomorphism f is an n -to-1 map from G onto $\text{im}(f)$.*

Proof This follows almost immediately from the previous proposition and Proposition 2.29. For any $h \in \text{im}(f)$ with $h = f(g)$ for some $g \in G$, we have

$$|f^{-1}(h)| = |gK| = |K| = n$$

so each element of $\text{im}(f)$ is mapped to by exactly n elements of G . $\square$

In the judgment of the most competent living mathematicians, Fräulein Noether was the most significant creative mathematical genius thus far produced since the higher education of women began. In the realm of algebra, in which the most gifted mathematicians have been busy for centuries, she discovered methods which have proved of enormous importance in the development of the present-day younger generation of mathematicians.

— Albert Einstein (1879–1955), Letter to the New York Times, 5 May 1935

4.3 The Isomorphism Theorems

IN THE PREVIOUS CHAPTER WE STUDIED NORMAL SUBGROUPS and used them to define quotient groups: new groups in which the normal subgroup (together with whatever it might happen to signify) is factored out and effectively annihilated or ignored. We saw how factoring a group G by its centre $Z(G)$ yielded a new group, and in a little while we will figure out what this particular quotient group tells us. We also saw how factoring by the commutator subgroup $[G, G]$ killed all the nonabelian structure in the group, leaving us with an abelianised version G^{ab} of the group.

Both the centre and the commutator subgroups are in some sense special: every group has one of each, and they're both well-defined (although in some cases they're either trivial or otherwise uninteresting).

Proposition 4.29 tells us that for any homomorphism $f: G \rightarrow H$, the kernel $\ker(f)$ is a normal subgroup of G . And since it's normal in G , we can go ahead and factor G by it to form the quotient group $G/\ker(f)$. Proposition 4.29 also tells us that any normal subgroup of G can be regarded as the kernel of some suitable homomorphism, so it's natural to ask what these quotient groups look like.

The answer is given by the following elegant theorem:

Theorem 4.40 (First Isomorphism Theorem) *Let $f: G \rightarrow H$ be a group homomorphism and let $K = \ker(f)$. Then the function $\phi: G/K \rightarrow \text{im}(f)$ given by $\phi(gK) = f(g)$ is an isomorphism. That is,*

$$G/\ker(f) \cong \text{im}(f).$$

Proof We need to show three things: that the function ϕ is well-defined, that it is a homomorphism, and that it is bijective.

Suppose, then, that $g_1, g_2 \in G$ and that the cosets g_1K and g_2K coincide; that is, $g_1K = g_2K$. Then by Proposition 2.26 it follows that $g_1 * g_2^{-1}$ lies in the kernel K . So $f(g_1 * g_2^{-1})$ lies in the image $f(K) = \{e_H\}$, and thus $f(g_1 * g_2^{-1}) = e_H$. Since f is a homomorphism,

$$f(g_1 * g_2^{-1}) = f(g_1) \bullet f(g_2^{-1}) = f(g_1) \bullet f(g_2)^{-1} = e_H$$

and therefore $f(g_1) = f(g_2)$. So if $g_1K = g_2K$ we have

$$\phi(g_1K) = f(g_1) = f(g_2) = \phi(g_2K)$$

and thus ϕ is well-defined.

The function ϕ is a homomorphism since

$$\begin{aligned} \phi((g_1K)(g_2K)) &= \phi((g_1 * g_2)K) \\ &= f(g_1 * g_2) \\ &= f(g_1) \bullet f(g_2) \\ &= \phi(g_1K)\phi(g_2K). \end{aligned}$$

It is clear that ϕ is surjective: any element of $\text{im}(f)$ is of the form $f(g)$ for some $g \in G$. And by the definition of ϕ , we have $f(g) = \phi(gK)$. Hence any element of $\text{im}(f)$ is of the form $\phi(gK)$ for some coset $gK \in G/K$.

Finally, we need to show that ϕ is injective. To see this, consider two elements $g_1, g_2 \in G$ such that $\phi(g_1K) = \phi(g_2K)$. Then $f(g_1) = f(g_2)$, which then means that

$$e_H = f(g_1) \bullet f(g_2)^{-1} = f(g_1) \bullet f(g_2^{-1}) = f(g_1 * g_2^{-1}).$$

This means that $g_1 * g_2^{-1} \in \ker(f)$, and by Proposition 2.26 this means that $g_1K = g_2K$ as cosets. Therefore ϕ is injective.

Hence $\phi: G/\ker(f) \rightarrow \text{im}(f)$ is a well-defined group isomorphism. $\square$

Let's look at some examples.

Example 4.41 In Example 3.27 we saw that factoring the Klein 4-group $V_4 = \{e, a, b, c\}$ by the subgroup $A = \{e, a\} \cong \mathbb{Z}_2$ yields a group which is itself isomorphic to $\mathbb{Z}_2$. We can rephrase this as a consequence of the First Isomorphism Theorem as follows. Let $f: V_4 \rightarrow \mathbb{Z}_2$ be the homomorphism which maps e and a to 0 and b and c to 1. Then $\ker(f) = A$ and f is surjective so $\text{im}(f) = \mathbb{Z}_2$. By the First Isomorphism Theorem we have

$$V_4/A = V_4/\ker(f) \cong \text{im}(f) = \mathbb{Z}_2$$

as expected.

Example 4.42 Revisiting Example 3.26 in the same way, we define a homomorphism $f: \mathbb{Z}_4 \rightarrow \mathbb{Z}_2$ such that $0, 2 \mapsto 0$ and $1, 3 \mapsto 1$. Then $\ker(f) = \{0, 2\}$ and $\text{im}(f) = \mathbb{Z}_2$, and so by the First Isomorphism Theorem we see that

$$\mathbb{Z}_4/\{0, 2\} = \mathbb{Z}_4/\ker(f) \cong \text{im}(f) = \mathbb{Z}_2$$

as expected.

The First Isomorphism Theorem is an important. In particular, it tells us we can study homomorphic images of a group G by examining its various quotient groups, since there is a well-defined correspondence between them. This correspondence is not always bijective, though, as different homomorphisms may have the same image:

Example 4.43 Let $V_4 = \{e, a, b, c\}$ be the Klein 4-group as usual. We define three group homomorphisms $\alpha, \beta, \gamma: V_4 \rightarrow \mathbb{Z}_2$ by

$$\alpha: \begin{cases} e, a \mapsto 0 \\ b, c \mapsto 1 \end{cases}, \quad \beta: \begin{cases} e, b \mapsto 0 \\ a, c \mapsto 1 \end{cases}, \quad \gamma: \begin{cases} e, c \mapsto 0 \\ a, b \mapsto 1 \end{cases}.$$

Let $A = \ker(\alpha)$, $B = \ker(\beta)$ and $C = \ker(\gamma)$. These three subgroups are obviously normal for a variety of reasons (they're kernels of group homomorphisms, V_4 is abelian, they're index-2 subgroups, etc) and lead to obviously isomorphic quotient groups

$$V_4/A = \{eA = aA, bA = cA\},$$

$$V_4/B = \{eB = bB, aB = cB\}$$

$$\text{and } V_4/C = \{eC = cC, aC = bC\}$$

which are in turn isomorphic to the images

$$\text{im}(\alpha) = \text{im}(\beta) = \text{im}(\gamma) = \mathbb{Z}_2.$$

We'll look more closely at the details of this correspondence later on. In particular, it will lead us into a discussion of group extensions and semidirect products. But for the moment, the moral of the First Isomorphism Theorem is that we can determine all homomorphic images of a group G from its various quotients.

This, of course, raises the question of why we might be interested in homomorphic images of a group when we could just be investigating the group itself. The answer is that with larger groups it can be quite time-consuming and difficult to study the group itself, whereas we can often find out quite a lot about the internal structure of a group from its homomorphic images, as the following examples demonstrate.

Example 4.44 Suppose G is a group of order 6 which maps surjectively onto the cyclic group $\mathbb{Z}_3$ via some homomorphism $f: G \rightarrow \mathbb{Z}_3$. Then we know that G must have a normal subgroup of order 2 since by Proposition 4.29 $K = \ker(f)$ is a normal subgroup of G , and Corollary 4.39 tells us that $|\ker(f)| = |f^{-1}(0)| = 2$. This means that $K \cong \mathbb{Z}_2$. It also means that G can only be the cyclic group $\mathbb{Z}_6$ since the only other group of order 6 is the dihedral group D_3 , which we already know doesn't have a normal subgroup of order 2.

In this case, merely knowing that such a homomorphism exists, without knowing anything else about it, was enough to help us figure out what the group was. This isn't always possible, but we can usually derive at least some information about the chosen group's internal structure, as the following example demonstrates.

Example 4.45 Suppose $f: G \rightarrow \mathbb{Z}_6$ is a surjective homomorphism from a group G of order 24 onto the cyclic group of order 6.

Then by Proposition 4.29 and Corollary 4.39 we know that G must have a normal subgroup of order 4 (which must be isomorphic to either $\mathbb{Z}_4$ or the Klein 4-group V_4) since $|\ker(f)| = |f^{-1}(0)| = 4$.

But we can do better than that. The image group $\mathbb{Z}_6$ has a normal subgroup $\{0, 3\}$ of order 2, so by Proposition 4.37 and Corollary 4.39, G has a normal subgroup of order $2 \times 4 = 8$.

Also, $\mathbb{Z}_6$ has a normal subgroup $\{0, 2, 4\}$ of order 3, so by the same argument we know G has a normal subgroup of order $3 \times 4 = 12$.

Up to isomorphism there are fifteen groups of order 24,² and we're still a little way from being able to say which one this is,³ but just knowing the existence of a surjective homomorphism onto $\mathbb{Z}_6$ has told us a surprising amount about its internal structure.

Lagrange's Theorem tells us that G might also contain subgroups of orders 2, 3 and 6, but doesn't guarantee their existence. However, Cauchy's Theorem helps us a little more, guaranteeing the existence of cyclic (albeit not necessarily normal) subgroups of orders 2 and 3.

For the next application of the First Isomorphism Theorem, we introduce three new constructs: the **normaliser** $N_G(H)$, the **centraliser** $Z_G(H)$ and the **inner automorphism group** $\text{Inn}(G)$.

1	$\mathbb{Z}_{24}$ $\cong \mathbb{Z}_8 \oplus \mathbb{Z}_3$
2	$\mathbb{Z}_{12} \oplus \mathbb{Z}_2$ $\cong \mathbb{Z}_6 \oplus \mathbb{Z}_4$ $\cong \mathbb{Z}_4 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_2$
3	$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6$ $\cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3$
4	D_{12}
5	$D_6 \times \mathbb{Z}_2 \cong D_3 \times \mathbb{Z}_2 \times \mathbb{Z}_2$
6	$D_4 \times \mathbb{Z}_3$
7	$D_3 \times \mathbb{Z}_4$ $\cong S_3 \times \mathbb{Z}_4$
8	S_4
9	$A_4 \times \mathbb{Z}_2$
10	$Q_8 \times \mathbb{Z}_3$
11	$SL_2(3)$
12	$\mathbb{Z}_3 \rtimes \mathbb{Z}_8$
13	$\mathbb{Z}_3 \rtimes D_4$ with kernel $\mathbb{Z}_2 \oplus \mathbb{Z}_2$
14	Dic_6
15	$Dic_3 \times \mathbb{Z}_2$

Table 4.1: The fifteen groups of order 24

² Just for interest, the fifteen groups of order 24 are shown in Table 4.1.

We've met the first eleven of these groups already, to a greater or lesser extent. (Recall that $SL_2(3)$ is the special linear group of 2×2 matrices with determinant 1 over the finite field $\mathbb{F}_3$.)

For the remaining four, Dic_n denotes the **dicyclic group** of order $4n$ and $\rtimes$ denotes the **semidirect product**. We'll meet both of these in later chapters.

³ It so happens that $SL_2(3)$ has no subgroups of order 12, so we can cross that one off the list of candidates for G .

Definition 4.46 Let S be a subset of a group G . The **normaliser** of S in G is the set

$$N_G(S) = \{g \in G : gSg^{-1} = S\}$$

where $gSg^{-1} = \{g * s * g^{-1} : s \in S\}$.

What this means is that the set gSg^{-1} consists of all possible conjugates of elements of S by some fixed element $g \in G$. So the normaliser $N_G(S)$ is the set of all elements $g \in G$ for which conjugation by g either leaves S unchanged or permutes its elements among themselves.

Example 4.47 Consider the subset $S = \{-1, 1\} \subset \mathbb{Z}$. For some arbitrary integer $n \in \mathbb{Z}$, the conjugate set is

$$n + S - n = \{n - 1 - n\} \cup \{n + 1 - n\} = \{-1, 1\} = S.$$

In other words, the set S is fixed by conjugation by any integer $n \in \mathbb{Z}$. What this means is that $N_{\mathbb{Z}}(S) = \mathbb{Z}$: the normaliser of S in $\mathbb{Z}$ is the whole of $\mathbb{Z}$.

Similarly, considering the subgroup $3\mathbb{Z} \leq \mathbb{Z}$, we see that

$$n + 3\mathbb{Z} - n = \{n + 3m - n : m \in \mathbb{Z}\} = \{3m : m \in \mathbb{Z}\} = 3\mathbb{Z}$$

and hence

$$N_{\mathbb{Z}}(3\mathbb{Z}) = \{n \in \mathbb{Z} : n + 3\mathbb{Z} - n = 3\mathbb{Z}\} = \{n \in \mathbb{Z} : 3\mathbb{Z} = 3\mathbb{Z}\} = \mathbb{Z}.$$

Thus $N_{\mathbb{Z}}(3\mathbb{Z}) = \mathbb{Z}$.

More generally, $N_G(S) = G$ for any abelian group G and any subset $S \subseteq G$. At first sight the normaliser doesn't look particularly useful: in the examples above it's just equal to the full group. But as usual, things become more interesting when we consider nonabelian groups.

Example 4.48 Let $G = D_3$ and consider the four subsets

$$S_1 = \{m_1, m_2, m_3\}, \quad S_2 = \{e, r, r^2\}, \quad S_3 = \{e, m_1\}, \quad S_4 = \{r\}.$$

Clearly the identity e is in $N_{D_3}(S)$ for any of these subsets, since $e * s * e^{-1} = s$ for any $s \in S$.

The normaliser $N_{D_3}(S_1)$ contains the rotations r and r^2 since

$$\begin{aligned} r * m_1 * r^2 &= m_2, & r^2 * m_1 * r &= m_3, \\ r * m_2 * r^2 &= m_3, & r^2 * m_2 * r &= m_1, \\ r * m_3 * r^2 &= m_1, & r^2 * m_3 * r &= m_2. \end{aligned}$$

Observe that not only is $g * s * g^{-1}$ in S_1 for each of these elements $g = r, r^2$, but that every element of S_1 is represented in this way. That is, gS_1g^{-1} is actually equal to S_1 rather than just being a subset of it.

The reflections m_1, m_2 and m_3 are in $N_{D_3}(S_1)$ as well, since

$$\begin{aligned} m_1 * m_1 * m_1 &= m_1, & m_2 * m_1 * m_2 &= m_3, & m_3 * m_1 * m_3 &= m_2, \\ m_1 * m_2 * m_1 &= m_3, & m_2 * m_2 * m_2 &= m_2, & m_3 * m_2 * m_3 &= m_1, \\ m_1 * m_3 * m_1 &= m_2, & m_2 * m_3 * m_2 &= m_1, & m_3 * m_3 * m_3 &= m_3. \end{aligned}$$

Hence $N_{D_3}(S_1) = D_3$.

The subset S_2 happens to be equal to the rotation subgroup R_3 , which we know is closed under conjugation by elements of D_3 since it's a normal subgroup of D_3 . A glance at the first three rows of Table 3.1 confirms that the normaliser $N_{D_3}(S_2) = D_3$.

Examining S_3 , we see that $m_1 \in N_{D_3}(S_3)$ since

$$m_1 * e * m_1 = e, \quad m_1 * m_1 * m_1 = m_1.$$

However, neither the rotations r and r^2 nor the other reflections m_1 and m_2 fix S_3 under conjugation, so $N_{D_3}(S_2) = \{e, m_1\}$.

Finally, the second row of Table 3.1 tells us that r is unchanged when conjugated with the identity e , r itself, or r^2 , but by no other elements. Hence $N_{D_3}(S_4) = \{e, r, r^2\}$.

This example displays a range of interesting behaviour. In all cases the normaliser $N_{D_3}(S)$ was a subgroup of D_3 , and as the following proposition shows, this wasn't a coincidence.

Proposition 4.49 *Let $S \subseteq G$ be a subset of some group G . Then the normaliser $N_G(S)$ is a subgroup of G .*

Proof The identity e is in $N_G(S)$ since $e*s*e^{-1} = e*s*e = s$ for all $s \in S$.

If $g_1, g_2 \in N_G(S)$ then their product $g_1 * g_2 \in N_G(S)$ too:

$$\begin{aligned} (g_1 * g_2)S(g_1 * g_2)^{-1} &= (g_1 * g_2)S(g_2^{-1} * g_1^{-1}) \\ &= g_1(g_2Sg_2^{-1})g_1^{-1} \\ &= g_1Sg_1^{-1} && \text{since } g_2 \in N_G(S) \\ &= S && \text{since } g_1 \in N_G(S). \end{aligned}$$

If $g \in N_G(S)$ then $S = gSg^{-1} = \{g*s*g^{-1} : s \in S\}$, and hence

$$Sg = \{s*g : s \in S\} = \{g*s : s \in S\} = gS$$

and therefore

$$g^{-1}Sg = \{g^{-1}*s*g : s \in S\} = S$$

so $g^{-1} \in N_G(S)$ as required. Thus $N_G(S)$ is a subgroup of G . $\square$

In Example 4.48 we considered two cases where the subset S happened to be a subgroup of the larger group D_3 . In the first of these cases, the subset S was actually the rotation subgroup $R_3 = \{e, r, r^2\}$, which

happens to be a normal subgroup of its normaliser $N_{D_3}(R_3) = D_3$. In the second case, the subgroup $\{e, m_1\}$ was its own normaliser. We call subgroups with this property (that is, subgroups $H \leq G$ such that $N_G(H) = H$) **self-normalising**. Clearly any self-normalising subgroup must be a normal subgroup of its own normaliser.

Example 4.50 Consider the following three subgroups of the symmetric group S_4 :

$$H_1 = \{\iota, (12)\},$$

$$H_2 = \{\iota, (123), (132)\},$$

$$\text{and } H_3 = \{\iota, (12)(34), (13)(24), (14)(23)\}$$

These subgroups have normalisers

$$N_{S_4}(H_1) = \{\iota, (12), (34), (12)(34)\} \cong V_4,$$

$$N_{S_4}(H_2) = \{\iota, (12), (23), (13), (123), (132)\} \cong S_3,$$

$$\text{and } N_{S_4}(H_3) = S_4$$

in S_4 , and normalisers

$$N_{A_4}(H_2) = \{\iota, (123), (132)\} \quad \text{and} \quad N_{A_4}(H_3) = A_4$$

in A_4 .

Again, we see the same phenomenon occurring:

$$H_1 \text{ is not normal in } S_4, \text{ but it is normal in } N_{S_4}(H_1) \cong V_4$$

$$H_2 \text{ is not normal in } S_4, \text{ but it is normal in } N_{S_4}(H_2) \cong S_3$$

$$H_3 \text{ is normal in both } S_4 \text{ and its normaliser } N_{S_4}(H_3) = S_4$$

$$H_2 \text{ is not normal in } A_4, \text{ but it is normal in } N_{A_4}(H_2) = H_2 \cong A_3$$

$$H_3 \text{ is normal in both } A_4 \text{ and its normaliser } N_{A_4}(H_3) = A_4$$

This happens in general: if $H \leq G$ is a subgroup of G then H is a normal subgroup of $N_G(H)$. Moreover, $N_G(H)$ is the largest subgroup of G in which H is normal, as the examples above show. When H is normal in G , its normaliser $N_G(H)$ is all of G , otherwise $N_G(H)$ is the largest subgroup with this property, in some cases just H itself.

A related concept is that of the **centraliser** of a subset of a group:

Definition 4.51 Let G be a group, and let S be a subset of G . Then the **centraliser** of S in G is the set

$$\begin{aligned} Z_G(S) &= \{g \in G : g * s * g^{-1} = s \text{ for all } s \in S\} \\ &= \{g \in G : g * s = s * g \text{ for all } s \in S\} \end{aligned}$$

This looks almost identical to the definition of the normaliser $N_G(S)$, but it's stricter: $N_G(S)$ consists of all elements of G that fix S *setwise* under conjugation (but might permute the elements of S) while $Z_G(S)$

consists of the elements of G that fix each *individual* element of S .

Equivalently, $Z_G(S)$ comprises those elements of G which commute with each element of S . The centraliser $Z_G(S)$ generalises the concept of the centre $Z(G)$ of a group,⁴ in the sense that $Z_G(G) = Z(G)$.

⁴ Definition 3.16, page 76.

Example 4.52 Let $G = D_3$ and let

$$S_1 = \{m_1, m_2, m_3\}, \quad S_2 = \{e, r, r^2\}, \quad S_3 = \{e, m_1\}, \quad S_4 = \{r\}.$$

Then

$$\begin{aligned} Z_{D_3}(S_1) &= \{e\}, & Z_{D_3}(S_2) &= \{e, r, r^2\}, \\ Z_{D_3}(S_3) &= \{e, m_1\}, & Z_{D_3}(S_4) &= \{e, r, r^2\}. \end{aligned}$$

The next example concerns subgroups of S_4 and A_4 .

Example 4.53 Let

$$\begin{aligned} H_1 &= \{\iota, (12)\}, \\ H_2 &= \{\iota, (123), (132)\}, \\ H_3 &= \{\iota, (12)(34), (13)(24), (14)(23)\}. \end{aligned}$$

Then these subgroups have centralisers

$$\begin{aligned} Z_{S_4}(H_1) &= \{\iota, (12), (34), (12)(34)\} \cong V_4, \\ Z_{S_4}(H_2) &= \{\iota, (123), (132)\} \cong \mathbb{Z}_3 \\ \text{and } Z_{S_4}(H_3) &= \{\iota, (12)(34), (13)(24), (14)(23)\} \cong V_4 \end{aligned}$$

in S_4 , and centralisers

$$\begin{aligned} Z_{A_4}(H_2) &= \{\iota, (123), (132)\} \cong \mathbb{Z}_3 \\ \text{and } Z_{A_4}(H_3) &= \{\iota, (12)(34), (13)(24), (14)(23)\} \cong V_4 \end{aligned}$$

in A_4 .

As these examples suggest, the centraliser of a subgroup is a subgroup:

Proposition 4.54 *Let G be a group, and let H be a subgroup of G . Then the centraliser $Z_G(H)$ is a subgroup of G .*

Proof The identity e lies in $Z_G(H)$ since $e * h = h * e$ for all $h \in H$.

If $g_1, g_2 \in Z_G(H)$ then $g_1 * g_2$ is also in $Z_G(H)$ since

$$(g_1 * g_2) * h = g_1 * h * g_2 = h * (g_1 * g_2).$$

If $g \in Z_G(H)$ then its inverse g^{-1} is also in $Z_G(H)$ since

$$g^{-1} * h = (h^{-1} * g)^{-1} = (g * h^{-1})^{-1} = h * g^{-1}$$

for all $h \in H$. Hence $Z_G(H)$ is a subgroup of G . $\square$

The third concept we need to introduce for our next application of the First Isomorphism Theorem is that of an automorphism group. Recall that an **automorphism** of a group G is an isomorphism $f: G \rightarrow G$. The automorphisms of a given group G themselves form a group:

Proposition 4.55 *Let G be a group, and let $\text{Aut}(G)$ be the set of all automorphisms $f: G \rightarrow G$. Then $\text{Aut}(G)$ forms a group under the usual composition operation for homomorphisms.*

Proof Composition of automorphisms is clearly associative, since composition of functions is always associative as long as the composite is actually defined. For an automorphism, the domain is the same as the codomain, so the composite will always be defined; it will also be a bijection (since the composite of two bijections is also a bijection) and must also be a homomorphism (since the composite of two homomorphisms is also a homomorphism). Therefore composition is a suitable associative binary operation $\circ: \text{Aut}(G) \times \text{Aut}(G) \rightarrow \text{Aut}(G)$. The identity map $\text{id}: G \rightarrow G$ is clearly an automorphism of G , and has the property that $f \circ \text{id} = f = \text{id} \circ f$ for any $f \in \text{Aut}(G)$. It thus acts as a suitable identity element in $\text{Aut}(G)$.

Every automorphism $f: G \rightarrow G$ has a well-defined, unique inverse $f^{-1}: G \rightarrow G$ such that

$$f \circ f^{-1} = \text{id} = f^{-1} \circ f.$$

Thus $\text{Aut}(G)$ is a group. □

As with permutation groups, we write the multiplication right-to-left rather than left-to-right in $\text{Aut}(G)$. As remarked earlier,⁵ there is a particular class of automorphisms that turn up in various circumstances:

Definition 4.56 Let G be a group. An **inner automorphism** of G is an automorphism of the form

$$\phi_g: G \rightarrow G; \quad h \mapsto g * h * g^{-1}$$

for some fixed $g \in G$ and all $h \in G$.

The inner automorphisms of a group form a group which we denote

$$\text{Inn}(G) = \{\phi_g : g \in G\},$$

and which is a subgroup of the full automorphism group $\text{Aut}(G)$.

We now have the ingredients we need for the following elegant result, an application of the First Isomorphism Theorem that describes the connection between normalisers, centralisers and automorphisms.

Theorem 4.57 (The Normaliser–Centraliser Theorem) *Let H be a subgroup of a group G . The quotient $N_G(H)/Z_G(H)$ is isomorphic to a subgroup of the automorphism group $\text{Aut}(H)$.*

Proof Let $f: N_G(H) \rightarrow \text{Aut}(H)$ be the map given by $f(g) = \phi_g$, the inner automorphism $\phi_g: H \rightarrow H$ such that $\phi_g(h) = g * h * g^{-1}$.

This map f is a homomorphism, since for any $g_1, g_2, h \in H$ we have

$$\phi_{g_1 * g_2}(h) = (g_1 * g_2) * h * (g_1 * g_2)^{-1}$$

⁵ Proposition 3.9, page 74.

$$\begin{aligned}
&= (g_1 * g_2) * h * (g_2^{-1} * g_1^{-1}) \\
&= g_1 * (g_2 * h * g_2^{-1}) * g_1^{-1} \\
&= \phi_{g_1}(\phi_{g_2}(h)) \\
&= (\phi_{g_1} \circ \phi_{g_2})(h)
\end{aligned}$$

and hence $f(g_1 * g_2) = f(g_1) \circ f(g_2)$ as required.

We want to apply the First Isomorphism Theorem to this homomorphism f , so we need to figure out what its kernel is. The kernel $\ker(f)$ will consist of all elements of $N_G(H)$ which map to the identity element in $\text{Aut}(H)$. This identity element is exactly the identity homomorphism $\text{id}: H \rightarrow H$. So we're looking for the elements of $N_G(H)$ that map to some inner automorphism ϕ_g that happens to be the identity.

In other words, we want the elements $g \in N_G(H)$ for which $\phi_g(h) = h$ for all elements $h \in H$. But these are exactly those $g \in N_G(H)$ for which $g * h * g^{-1} = h$ and hence $\ker(f) = Z_G(H)$, the centraliser of H in G .

By the First Isomorphism Theorem, then,

$$N_G(H) / \ker(f) = N_G(H) / Z_G(H) \cong \text{im}(f) \leq \text{Aut}(H)$$

as required. $\square$

We can now use this result to tell us what it means if we take the quotient of a group G by its centre $Z(G)$: we get the inner automorphism group of G .

Corollary 4.58 *Let G be a group. Then $G/Z(G) \cong \text{Inn}(G)$.*

Proof The normaliser $N_G(G)$ is G itself, and the centraliser $Z_G(G)$ is, as remarked earlier, just the centre $Z(G)$ of G . The image of the homomorphism $f: N_G(G) \rightarrow \text{Aut}(G)$ is exactly the subgroup $\text{Inn}(G)$ of inner automorphisms of G . Hence, by the Normaliser–Centraliser Theorem, $G/Z(G) \cong \text{Inn}(G)$. $\square$

Now recall⁶ that if G is a group, and H and K are both subgroups of G , we denote by HK the set

$$HK = \{h * k : h \in H, k \in K\}.$$

Soon we'll study and prove the Second Isomorphism Theorem, but to provide a little motivation we'll look at a couple of examples: one involving the quaternion group Q_8 and the other involving the symmetric group S_4 and the alternating group A_4 .

Example 4.59 Let $G = Q_8$, the quaternion group we met in Example 1.42, and let $H = \{\pm E, \pm I\}$ and $K = \{\pm E, \pm J\}$. Then $HK = Q_8 = G$ and $H \cap K = \{\pm E\} \cong \mathbb{Z}_2$. (Proposition 2.9 says

⁶ Definition 2.10, page 45.

that $H \cap K$ must be a subgroup of G , but here it's fairly obvious.) All of H , K and $H \cap K$ are normal subgroups of $HK = G = Q_8$, so we can use them to form quotient groups. From Example 2.21 we can see that $G/H \cong \mathbb{Z}_2$. Also, $H \cap K = \{\pm E\} \cong \mathbb{Z}_2$ and $H \cong K \cong V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. So, by Example 3.27 we can see that $H/(H \cap K) \cong \mathbb{Z}_2$. Hence $HK/H \cong H/(H \cap K)$.

Example 4.60 Let $G = S_4$, let $K = A_4$, and let $H = \{I, (123), (132)\}$. Then $HK = A_4$ and $H \cap K = H$, so both HK/K and $H/(H \cap K)$ are trivial, and hence isomorphic to each other.

The behaviour exhibited in these examples, that $HK/K \cong H/(H \cap K)$, is true in general. More than that, it turns out that we only really need one of H and K to be a normal subgroup of G , but in that case we need the following lemma, which is a generalisation of Proposition 2.9.

Lemma 4.61 *Let G be a group, let H be a subgroup of G , and let N be a normal subgroup of G . Then $H \cap N$ is a normal subgroup of H .*

Proof Proposition 2.9 tells us that the intersection $H \cap N$ is a subgroup of G , and since it's clearly a subset of H it must also be a subgroup of H . So all that remains is to show that it's a normal subgroup of H .

Let $h \in H$ and $k \in H \cap N$. We need to show that the conjugate $h * k * h^{-1}$ also belongs to $H \cap N$. It certainly belongs to H , since H is closed under the multiplication operation $*$ inherited from G , and it must also belong to N since N is a normal subgroup of G and hence closed under conjugation. Therefore $h * k * h^{-1}$ must belong to the intersection $H \cap N$, and thus $H \cap N$ is a normal subgroup of H . $\square$

We're now ready to state and prove the Second Isomorphism Theorem:

Theorem 4.62 (Second Isomorphism Theorem) *Let G be a group, let H be a (not necessarily normal) subgroup of G , and let N be a normal subgroup of G . Then*

$$HN/N \cong H/(H \cap N).$$

Proof Lemma 4.61 tells us that $H \cap N$ is a normal subgroup of H , so the statement makes sense. We can use the First Isomorphism Theorem⁷ to prove this result; the key is to construct the right surjective homomorphism $f: HN \rightarrow H/(H \cap N)$. If we can find such a homomorphism that happens to have $\ker(f) = N$, then the Second Isomorphism Theorem will follow immediately.

A typical element of HN is of the form $h * n$ where $h \in H$ and $n \in N$, while a typical element of the quotient $H/(H \cap N)$ is a coset $k(H \cap N)$

⁷Theorem 4.40, page 117.

for some $k \in H$.

With this in mind, we define a function $f: HN \rightarrow H/(H \cap N)$ mapping the element $h * n \in HN$ to the coset $h(H \cap N)$ in $H/(H \cap N)$.

We have to show four things: that f is well-defined, that it's surjective onto $H/(H \cap N)$, that it's a group homomorphism, and that its kernel is the normal subgroup N .

Dealing with the first of these, let $h_1, h_2 \in H$ and $n_1, n_2 \in N$ such that

$$h_1 * n_1 = h_2 * n_2$$

in HN . Rearranging this, we get

$$h_2^{-1} * h_1 = n_2 * n_1^{-1}.$$

On the left-hand side, $h_2^{-1} * h_1$ is clearly an element of H , while on the right-hand side, $n_2 * n_1^{-1}$ is clearly an element of N . Since they're equal, it must be the case that both belong to H and N . So both sides, and specifically $h_2^{-1} * h_1$, are in the intersection $H \cap N$. Hence by Proposition 2.26, since $h_2^{-1} * h_1 \in H \cap N$, the cosets $h_1(H \cap N)$ and $h_2(H \cap N)$ are equal, and so $f(h_1 * n_1) = h_1(H \cap N) = h_2(H \cap N) = f(h_2 * n_2)$. Thus f is well-defined.

To show that f is a group homomorphism, we need the following observation: given any $h \in H$ and $n \in N$ there exists some element $k \in N$ such that $h * n = k * h$. This is because N , being a normal subgroup, is closed under conjugation and so there is some element $k \in N$ such that $h * n * h^{-1} = k$. Rearranging this, we get the desired fact. Then, for any $h_1, h_2 \in H$ and $n_1, n_2 \in N$ we have

$$\begin{aligned} f((h_1 * n_1) * (h_2 * n_2)) &= f(h_1 * (n_1 * h_2) * n_2) \\ &= f(h_1 * h_2 * n_3 * n_2) \quad (\text{for } n_3 = h_2^{-1} * n_1 * h_2) \\ &= (h_1 * h_2)(H \cap N) \\ &= h_1(H \cap N) \cdot h_2(H \cap N) \\ &= f(h_1 * n_1) \cdot f(h_2 * n_2) \end{aligned}$$

and hence f is a homomorphism.

To see that f is surjective, consider any $h \in H$ and form the coset $h(H \cap N)$. We need to find some element $h * n$ of HN which maps to this coset. Setting $n = e \in N$ to get the product $h * e \in HN$ suffices admirably, since $f(h * e) = h(H \cap N)$ as required.

Finally, we need to show that $\ker(f) = N$. This kernel consists of all elements $h * n \in HN$ such that $f(h * n) = H \cap N$. From this, we see that the only elements $h \in H$ for which $h * n \mapsto H \cap N$ are those satisfying $h(H \cap N) = H \cap N$, and Proposition 2.26 tells us that these are exactly the elements of H which also belong to N . So $\ker(f) = (H \cap N)N$. But this is exactly the same as N , so $\ker(f) = N$.



Wikimedia Commons

Emmy Noether (1882–1935) made important contributions to algebra and physics: as well as the three Isomorphism Theorems as we know them today, she is also remembered for her work on ring theory and for Noether's Theorem, which describes the connection between symmetry and conservation laws in physics.

Born in 1882 to a Jewish family in Erlangen, Bavaria, she was the daughter of the algebraic geometer Max Noether and his wife Ida Kauffmann. She displayed a talent for languages and qualified as a teacher of French and English before deciding to study mathematics at university. At the time, women were only permitted to study unofficially at German universities, but she successfully persuaded the University of Erlangen to admit her, and completed her doctoral thesis *Über die Bildung des Formensystems der ternären biquadratischen Form* in 1907. She taught without pay at Erlangen until David Hilbert and Felix Klein invited her to Göttingen in 1915, where she taught (initially under Hilbert's name) until the early 1930s.

In 1933 she was dismissed from Göttingen due to laws passed by the Nazi government forbidding Jewish academics from holding posts at German universities. She was promptly offered a visiting professorship at Bryn Mawr College in Pennsylvania, and worked there and at the Institute for Advanced Study in Princeton, New Jersey, until her death in 1935.

By the First Isomorphism Theorem, then, we have

$$HN/N = HN/\ker(f) \cong \text{im}(f) = H/(H \cap N)$$

as required. $\square$

The next example concerns a nested sequence of normal subgroups of S_4 , and will lead to the last big theorem in this chapter.

Example 4.63 Let

$$G = S_4, \quad H = A_4 \quad \text{and} \quad K = \{1, (12)(34), (13)(24), (14)(23)\}.$$

Then H and K are both normal subgroups of G , and K is a normal subgroup of H , so we can form the quotient groups

$$G/H \cong \mathbb{Z}_2, \quad G/K \cong S_3, \quad H/K \cong \mathbb{Z}_3.$$

The quotient H/K is isomorphic to an index-2 (and hence normal) subgroup of G/K , so we can form the quotient group $(G/K)/(H/K)$, which happens to be isomorphic to $\mathbb{Z}_2$ and therefore also to G/H :

$$G/H \cong (G/K)/(H/K)$$

It is not immediately obvious that (H/K) is normal in (G/K) , and therefore that the quotient $(G/K)/(H/K)$ makes sense in general. Nevertheless, this is indeed the case, as the next lemma shows.

Lemma 4.64 *If G is a group, and H and K are normal subgroups of G such that K is also a normal subgroup of H , then the quotient H/K is also a normal subgroup of the quotient G/K .*

Proof We need to show that H/K is closed under conjugation by elements of G/K . The elements of H/K are cosets of the form hK for some element $h \in H$, and the elements of G/K are cosets of the form gK for some element $g \in G$.

Consider the conjugate $(gK)(hK)(gK)^{-1}$. This is equal to

$$(gK)(hK)(g^{-1}K) = ((g * h)K)(g^{-1}K) = (g * h * g^{-1})K.$$

But since H is a normal subgroup of G , the conjugate $g * h * g^{-1}$ lies in H and hence the coset $(g * h * g^{-1})K$ is in the quotient H/K as required, and therefore H/K is a normal subgroup of G/K . $\square$

Having proved this, we can now state and prove the Third Isomorphism Theorem.

Theorem 4.65 (Third Isomorphism Theorem) *Let G be a group, and let H and K be normal subgroups of G such that K is also a normal subgroup of H . Then*

$$G/H \cong (G/K)/(H/K).$$

Proof As with the Second Isomorphism Theorem, we apply the First Isomorphism Theorem to a suitable group homomorphism.

The elements of $(G/K)/(H/K)$ are cosets of H/K determined by elements of G/K , which are themselves cosets of K determined by elements of G . So an element of G/K will be a coset of the form gK , where $g \in G$, and hence an element of $(G/K)/(H/K)$ will be a coset of the form $(gK)(H/K)$. With this in mind, define

$$f: G \rightarrow (G/K)/(H/K); \quad g \mapsto (gK)(H/K).$$

This map f is well-defined, but we still need to show that it's a homomorphism, that it's surjective onto $(G/K)/(H/K)$, and that its kernel is H .

Suppose that $g_1, g_2 \in G$. Then

$$\begin{aligned} f(g_1 * g_2) &= ((g_1 * g_2)K)(H/K) \\ &= ((g_1K)(g_2K))(H/K) \\ &= ((g_1K)(H/K))((g_2K)(H/K)) \\ &= f(g_1)f(g_2) \end{aligned}$$

and thus f is a homomorphism.

This homomorphism f is clearly surjective. Let $(gK)(H/K)$ be some element of $(G/K)/(H/K)$. Then $g \in G$ is mapped to this coset by f .

Finally, the kernel $\ker(f)$ consists of those elements of G which are mapped to the coset H/K in $(G/K)/(H/K)$. These are those elements $g \in G$ for which $gK \in H/K$. But the elements satisfying this condition are exactly the elements of H . So $\ker(f) = H$, and by the First Isomorphism Theorem we have

$$G/H = G/\ker(f) \cong \text{im}(f) = (G/K)/(H/K)$$

as required. $\square$

The Third Isomorphism Theorem says that if we have three groups $K \trianglelefteq H \trianglelefteq G$, then we get the same result if we factor G/H as we do if we factor both G and H by K and then factor the resulting quotients: we can either collapse G to G/H in one step, or go via the intermediate factor-by- K step, and in both cases we get the same answer.

The three Isomorphism Theorems are generally attributed to the German mathematician Emmy Noether (1882–1935), who stated them in a little more generality (and using different notation) than this in her important 1927 paper *Abstrakter Aufbau der Idealtheorie in algebraischen Zahl- und Funktionenkörpern* (Abstract composition of ideal theory in algebraic number and function fields),⁸ although some of the ideas were present in her earlier work. Hints can also be seen in the work of Richard Dedekind (1831–1916) and in the classic text *Traité des substitutions et des équations algébriques* (Treatise on permutations and algebraic equations)⁹ by Camille Jordan (1838–1922).

⁸ E Noether, *Abstrakter Aufbau der Idealtheorie in algebraischen Zahl- und Funktionenkörpern*, *Mathematische Annalen* 96 (1927) 26–61.

⁹ C Jordan, *Traité des substitutions et des équations algébriques*, Gauthier-Villars (1870)

Summary

¹⁰ Definition 4.1, page 104.

¹¹ Definition 1.17, page 11.

In this chapter we studied **homomorphisms**,¹⁰ functions $f: G \rightarrow H$ from one group to another that preserve some aspects of the group structure, in the sense that $f(g_1 * g_2) = f(g_1) \bullet f(g_2)$ for any elements $g_1, g_2 \in G$. We've already met the concept of a bijective homomorphism, an **isomorphism**,¹¹ and in this chapter we also saw several examples of injective homomorphisms or **monomorphisms** (sometimes denoted with a tailed arrow $f: G \rightarrowtail H$) and surjective homomorphisms or **epimorphisms** (sometimes denoted with a two-headed arrow $f: G \twoheadrightarrow H$). In particular, the **identity** isomorphism $\text{id}_G: G \rightarrow G$ is always defined for any group G , but more generally an isomorphism from a group to itself is called an **automorphism**, and a non-bijective homomorphism from a group to itself is called an **endomorphism**. For any groups G and H there is a unique **zero** homomorphism $z: G \rightarrow H$ which maps every element of G to the identity element e_H in H .

¹² Proposition 4.11, page 105.

Reassuringly, composing two homomorphisms yields another homomorphism.¹²

¹³ Example 4.14, page 106.

¹⁴ Example 4.16, page 106.

¹⁵ Example 4.15, page 106.

Important examples of monomorphisms are given by **inclusion** homomorphisms $i: H \hookrightarrow G$ mapping a subgroup $H \leq G$ into a larger group.^{13,14} Important examples of epimorphisms include **projection** homomorphisms.¹⁵ Examples of both can often be constructed by considering the particular context of the groups in question: in the case of dihedral or other isometry groups, we can sometimes use our geometric intuition to define homomorphisms.¹⁶

¹⁶ Example 4.17, page 107.

¹⁷ Example 4.18, page 107.

Another class of homomorphisms derive from the quotient groups we learned about in the last chapter: given any normal subgroup H of a group G we can define a canonical, surjective **quotient homomorphism** $q: G \rightarrow G/H$.¹⁷ In particular, factoring a group G by its commutator subgroup $[G, G]$ yields the **abelianisation homomorphism** $\alpha: G \rightarrow G^{\text{ab}} = G/[G, G]$.¹⁸ These quotient homomorphisms, in particular the abelianisation homomorphism, have an interesting and useful property: given a homomorphism $f: G \rightarrow H$ there is a well-defined **induced homomorphism** $f^{\text{ab}}: G^{\text{ab}} \rightarrow H^{\text{ab}}$.^{19,20}

¹⁸ Example 4.21, page 108.

¹⁹ Proposition 4.22, page 108.

²⁰ Example 4.23, page 109.

²¹ Proposition 4.24, page 110.

Homomorphisms map identity elements to identity elements, and inverses to inverses, in the sense that if $f: G \rightarrow H$ is a homomorphism, then $f(e_G) = e_H$ and $f(g^{-1}) = f(g)^{-1}$ for any $g \in G$.²¹

²² Proposition 4.25, page 110.

²³ Example 4.26, page 111.

Since homomorphisms are functions, we can discuss their images; we denote the **image** of a homomorphism $f: G \rightarrow H$ by $\text{im}(f)$. It turns out that the image of a homomorphism is a subgroup of its codomain,²² but in general not always a normal subgroup.²³ If a

homomorphism $f: G \rightarrow H$ is injective, then its image $\text{im}(f)$ is isomorphic to the domain G of f .²⁴

²⁴ Example 4.27, page 111.

Certain problems in linear algebra reduce to studying the kernel of some linear map, and with this in mind we introduced the concept of the **kernel** $\ker(f)$ of a group homomorphism $f: G \rightarrow H$, which comprises all elements of the domain G which map to the identity e_H in the codomain H .²⁵ The kernel of f must at least contain the identity element e_G .²⁶ The kernel is a subgroup of the domain just as the image is a subgroup of the codomain. More than that, however, the kernel is a *normal* subgroup, and all normal subgroups of a given group G can be characterised as the kernel of some suitable homomorphism defined on G .²⁷ As noted earlier, the kernel must contain the identity e_G ; if, however, the kernel *only* contains the identity, then this is the same as saying that the homomorphism in question is injective.²⁸

²⁵ Definition 4.28, page 112.

²⁶ Proposition 4.24, page 110.

²⁷ Proposition 4.29, page 113.

²⁸ Proposition 4.35, page 114.

Generalising the definition of the kernel we arrived at the concept of the **inverse image** or **preimage**.²⁹ We found that the inverse image of a subgroup of the codomain is itself a subgroup of the domain, and the inverse image of a normal subgroup is also a normal subgroup.³⁰ In fact, for any homomorphism $f: G \rightarrow H$, the inverse image $f^{-1}(h)$ of a single element h of the codomain H will either be the empty set $\emptyset$ (if $h \notin \text{im}(f)$) or a coset of the kernel $\ker(f) \trianglelefteq G$.³¹ Putting this together with Proposition 2.29 we learned that if $|\ker(f)| = n$, then f is an n -to-1 map from G onto $\text{im}(f) \leq H$.³²

²⁹ Definition 4.36, page 115.

³⁰ Proposition 4.37, page 115.

³¹ Proposition 4.38, page 115.

³² Corollary 4.39, page 116.

The remainder of the chapter was devoted to the three Isomorphism Theorems. The first of these was motivated by the observation³³ that any normal subgroup can be regarded as the kernel of some appropriate homomorphism, and we wondered what relation the corresponding quotient group might have to the homomorphism in question. The answer was given by the First Isomorphism Theorem:³⁴ for any homomorphism $f: G \rightarrow H$, the quotient group $G/\ker(f)$ is isomorphic to the image subgroup $\text{im}(f)$ in H .

³³ Proposition 4.29, page 113.

³⁴ Theorem 4.40, page 117.

After a short discussion of the usefulness of studying homomorphic images of groups, and an investigation of some of the structural information we can recover from them, we looked at various applications of the First Isomorphism Theorem. One of the most important of these was the Normaliser–Centraliser Theorem,³⁵ which as a corollary³⁶ gave us an elegant and concise interpretation of what we get when we factor a group G by its centre $Z(G)$. During this discussion we introduced four new concepts: the **normaliser** $N_G(S)$ of a subset $S \subseteq G$ of a group G ,³⁷ the **centraliser** $Z_G(S)$ of a subset $S \subseteq G$ in G , the **automorphism group** $\text{Aut}(G)$ ³⁸ and the subgroup $\text{Inn}(G)$ of **inner automorphisms**.³⁹

³⁵ Theorem 4.57, page 124.

³⁶ Corollary 4.58, page 125.

³⁷ Definition 4.46, page 120.

³⁸ Proposition 4.55, page 124.

³⁹ Definition 4.56, page 124.

⁴⁰ Proposition 4.49, page 121.

⁴¹ Exercise 4.1, page 132.

⁴² Definition 3.16, page 76.

The normaliser $N_G(S)$ is always a subgroup of G^{40} , and in the case where the subset S is a *subgroup* of G it so happens that S is a normal subgroup of $N_G(S)$; in fact, we can regard $N_G(S)$ as the largest subgroup of G in which S is normal.⁴¹ The centraliser $Z_G(S)$ is a generalisation of the concept of the **centre** $Z(G)$ of a group G ,⁴² and also a slightly stricter object than the normaliser $N_G(S)$. It turns out that $Z_G(S)$ is always a normal subgroup of $N_G(S)$, and the Normaliser–Centraliser Theorem tells us that if H is a subgroup of G , then the quotient $N_G(H)/Z_G(H)$ is isomorphic to a subgroup of the automorphism group $\text{Aut}(H)$. Considering the special case where $H = G$ we find that the centraliser $Z_G(G)$ is just the centre $Z(G)$, and the normaliser $N_G(G)$ is G itself; this leads to the discovery that the quotient $G/Z(G)$ is isomorphic to the inner automorphism group $\text{Inn}(G)$.

⁴³ Theorem 4.62, page 126.

The Second Isomorphism Theorem⁴³ can be proved by applying the First Isomorphism Theorem to a suitable homomorphism. Suppose that G is a group, H is a subgroup of G and N is a normal subgroup of G . A short lemma confirms that the intersection $H \cap N$ is not just a subgroup of G ,⁴⁴ but also a normal subgroup of H .⁴⁵ The Second Isomorphism Theorem states that the quotient HN/N (where HN denotes the subgroup consisting of all products of an element of H with an element of N) is isomorphic to the quotient $H/(H \cap N)$.

⁴⁴ Proposition 2.9, page 45.

⁴⁵ Lemma 4.61, page 126.

⁴⁶ Theorem 4.65, page 128.

The Third Isomorphism Theorem⁴⁶ can also be proved by applying the First Isomorphism Theorem to an appropriate homomorphism. It states that if we have three groups G , H and K such that H and K are normal subgroups of G , and K is also a normal subgroup of H , then the quotient $(G/K)/(H/K)$ is isomorphic to the quotient G/H .

References and further reading

Some details on Emmy Noether's work on the isomorphism theorems (and her contributions to algebraic topology) can be found in the following article:

- C McLarty, *Emmy Noether's 'set-theoretic' topology: from Dedekind to the rise of functors*, in: *The Architecture of Modern Mathematics: Essays in history and philosophy*, ed. by J Gray and J Ferreirós, Oxford University Press (2006) 211–235

Exercises

- 4.1 Show that if H is a subgroup of G then H is a normal subgroup of $N_G(H)$. Show also that $N_G(H)$ is the largest subgroup of G in which H is normal.
- 4.2 Let G be a group. Show that the set $\text{Inn}(G)$ of inner automorphisms of G is a group. Show that it is a subgroup of $\text{Aut}(G)$.

5 Presentations

I am not yet so lost in lexicography, as to forget that *words are the daughters of the earth, and that things are the sons of heaven*. Language is only the instrument of science, and words are but the signs of ideas: I wish, however, that the instrument might be less apt to decay, and that signs might be permanent, like the things which they denote.

— Samuel Johnson (1709–1784),
preface to *A Dictionary of the English Language* (1755)

IN CHAPTER 1 we met the concept of a **generator** of a group.¹ We said that if $\{g_1, g_2, \dots\}$ is a (possibly infinite) collection of elements of a group G , and if any element of G can be expressed as a finite product of some or all of these elements and their inverses, then G is **generated** by $g_1, g_2, \dots$, which we call **generators** of G .

¹ Definition 1.22, page 12.

In the case of an infinite cyclic group $\mathbb{Z}$, we can generate the entire group by just using the element $1 \in \mathbb{Z}$ and its inverse -1 . Or, if we're using multiplicative notation, we can consider the group $\{\dots, t^{-2}, t^{-1}, 1, t, t^2, \dots\}$ of integer powers of some symbol t . This is clearly isomorphic to $\mathbb{Z}$ by the mapping $t^k \mapsto k \in \mathbb{Z}$. So, we can represent $\mathbb{Z}$ by $\langle 1 \rangle$, the group generated additively by the single element 1, or $\langle t \rangle$, the group generated multiplicatively by the single element t . (In the process we'll have to be careful to say whether we're using additive or multiplicative notation unless it's obvious from context.)

For the finite cyclic group $\mathbb{Z}_n$ we can also generate the entire group by a single element $1 \in \mathbb{Z}_n$. Or, using multiplicative notation again, we can consider the 0th to $(n-1)$ th powers of some single element t , to get $\{1, t, t^2, \dots, t^{n-1}\}$. This latter group is obviously isomorphic to $\mathbb{Z}_n$ by the same mapping as before: $t^k \mapsto k \in \mathbb{Z}_n$. Again, we can denote $\mathbb{Z}_n$ by $\langle 1 \rangle$ or $\langle t \rangle$. Or, for that matter, by $\langle k \rangle$ or $\langle t^k \rangle$ for any $k \in \mathbb{Z}_n$ such that $\gcd(k, n) = 1$.²

² Proposition 1.26, page 13.

We called groups which can be generated by a single element **cyclic** groups.³ But we can construct other groups as products of small numbers of generators too. The dihedral group D_n can be reconstructed by just using two of its elements: the basic $\frac{2\pi}{n}$ rotation element r and the reflection m_1 . The Klein group $V_4 = \{e, a, b, c\}$ can be generated by two elements as well: a and b .

³ Definition 1.23, page 12.

We call groups that can be constructed in this way by a finite number of their elements **finitely generated**. Groups which require an infinite set of elements are said to be **infinitely generated**; we'll see later that the additive group $\mathbb{Q}$ of rational numbers is infinitely generated.

In this chapter, we're going to study how to construct groups from (finite or infinite) sets of generators. As is so often the case, the details are a bit more complicated than the above examples suggest. In particular, we've glossed over a subtle but important point: it's all very well giving a set of generating elements for a group, but we also need to specify how those generators interact with each other, and what their orders are. It's not enough to say that $\mathbb{Z}_n$ is generated by a single generator t , we need also to specify that $t^n = 1$. We can write this compactly as

$$\mathbb{Z}_n \cong \langle t : t^n = 1 \rangle$$

which we interpret as the group generated by a single generator t , subject to the constraint or **relation** $t^n = 1$. We call an expression of this type a **presentation** for $\mathbb{Z}_n$, and will develop this idea in more detail over the rest of this chapter.

I hope we shall make everything as plain and as simple to you as we can. I would never use a long word, even, where a short one would answer the purpose.

— Oliver Wendell Holmes Sr
(1809–1894),

Medical Essays, 1842–1882 (1891) 302

5.1 Free groups

THE INFINITE CYCLIC GROUP $\mathbb{Z}$ is the most general possible group we can construct from a single generator, and as we saw just now, we can make the finite cyclic group $\mathbb{Z}_n$ by taking $\mathbb{Z}$ and imposing a constraint on it. Or, to put it another way, $\mathbb{Z}_n$ is the same as $\mathbb{Z}$ except that we replace all occurrences of t^n by 1. Amongst other things, this means that $t^{-1} = t^{n-1}$, that $t^{-2} = t^{n-2}$ and so on. We want to devise a similar way of constructing non-cyclic groups like D_n and V_4 , and to do this we first have to generalise our idea of $\mathbb{Z}$ as the most general possible group constructible from a single generator: we want to know how to construct the most general possible group with two or more generators.

Suppose that we have two symbolic variables x and y . At present these are just symbols: don't worry about whether they're numbers, symmetry operations, permutations or anything else. For the moment, we're just going to manipulate them in a purely formal and abstract way.

We can form finite **strings** of these two symbols, like $xxxyxyyx$, yy , $xyxy$ and x . We call strings like this **words** in x and y ; we should also allow the trivial case of the **empty word** which we'll usually denote e (although some books use other symbols like ι , a pair of empty parentheses $()$, 0 or 1).

We can simplify words by collapsing together repeated adjacent symbols of the same type and using integer exponents: the first two

examples become, respectively, x^3yxy^2x and y^2 . We don't lose any information by doing this: if we really wanted to, we could re-expand the collapsed words to get the original forms.

What we're not allowed to do is to arbitrarily reorder the symbols in a given word. So although $x^2y^3 = xxyyyy$, neither are equal to $yxyxy$ or y^3x^2 , for example.

We can, if we so choose, interpolate, prepend or append any number of copies of the empty word e , so $x^2y^3 = xxyyyy = xexyyeye = xexy^2eye$.

Typically, we'll **reduce** words by deleting any occurrences of e (except in the case of the empty word e itself) and combining repeated adjacent instances of the same symbols. Any word thus has a unique simplest reduced form.

Given two words, we can **concatenate** them to get a new word. For example, $xyx \cdot y^3x = xyxy^3x$. This might in some cases result in repeated adjacent symbols of the same type, as with $xy^2 \cdot y^4x^5 = xy^2y^4x^5$ so we follow such a concatenation with any necessary reduction operations in order to end up with a word in its simplest form.

It's fairly obvious that this concatenation–reduction operation is associative: given three words w_1 , w_2 and w_3 , the concatenation $(w_1 \cdot w_2) \cdot w_3$ gives the same word (when all necessary reductions have taken place) as the concatenation $w_1 \cdot (w_2 \cdot w_3)$.

The empty word serves as an identity with respect to this concatenation operation, in the sense that for any word w formed from the symbols x and y it's always the case that

$$e \cdot w = w = w \cdot e.$$

If we let W be the set of all finite, reduced words in the two symbols x and y , then the concatenation–reduction operation is an associative binary operation defined on this set, and the empty word e behaves like an identity. What we've just shown is the following:

Proposition 5.1 *Let W be the set of finite, reduced words in two symbols x and y , and let $\cdot : W \times W \rightarrow W$ be the concatenation–reduction operation. Then $(W, \cdot)$ is a monoid.*

In fact, this is the most general possible monoid we can get from two generating symbols. Monoids are all very well, and as remarked in Chapter 1 they have numerous applications in parts of mathematics and computer science, but at the moment we're primarily interested in groups. Is there some way we can modify or extend this construction to get the most general possible *group* generated by two symbols x and y ? As it happens, yes there is, and the key is to think about what's missing from a monoid that stops it being a group, namely inverses.

So, back to our symbols x and y . Introduce two more symbols x^{-1} and y^{-1} . These will function as the inverses of, respectively, x and y , but at the moment we're still going to treat them as formal symbols.

Just as before, we'll consider finite words in these four symbols, but we need to slightly extend our idea of reduction. In addition to combining repeated, adjacent symbols of the same type using positive integer exponents, we'll also combine repetitions of the new symbols x^{-1} and y^{-1} using negative integer exponents. Additionally, we'll cancel adjacent pairs of x and x^{-1} , and adjacent pairs of y and y^{-1} .

For example, the word $xx^{-1}x^{-1}x^{-1}yxy^{-1}y$ reduces to $xx^{-1}x^{-1}y y x$, then to $x^{-1}y y x$ and finally to $x^{-1}y^2x$, while the word $y^{-1}y^{-1}y^{-1}y x x x$ reduces to $y^{-2}x^3$ in a similar way.

Again, we don't allow arbitrary reordering, just cancellation and grouping of adjacent symbols.

So, let W be the set of all finite, reduced words in the symbols x , y , x^{-1} and y^{-1} , and let $\cdot : W \times W \rightarrow W$ be the concatenation–reduction operation; this is a well-defined, associative binary operation on the set W . As before, the empty word e serves as an identity.

All that remains is to ensure the existence of inverses in W . We know that x and y have inverses, because we introduced x^{-1} and y^{-1} for exactly this purpose. But we need to do a little more work to ensure that any word in W has a well-defined inverse. The clue is given by the discussion of inverses of products on page 9, where we deduced that $(g * h)^{-1} = h^{-1} * g^{-1}$ for any two elements g and h of some group G . For any word $w = x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k}$ where $x_{i_1}, \dots, x_{i_k} \in \{x, y\}$ and $n_1, \dots, n_k \in \mathbb{Z}$, the inverse of w is

$$w^{-1} = x_{i_k}^{-n_k} \dots x_{i_1}^{-n_1}.$$

To see this, concatenate w and w^{-1} either way round, reduce, and see that you end up with the empty word e :

$$\begin{aligned} w \cdot w^{-1} &= x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k} \cdot x_{i_k}^{-n_k} \dots x_{i_1}^{-n_1} = e \\ w^{-1} \cdot w &= x_{i_k}^{-n_k} \dots x_{i_1}^{-n_1} = e \cdot x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k} \end{aligned}$$

We now have all the ingredients to make a group.

Proposition 5.2 *Let F_2 be the set of finite, reduced words in the symbols x , y , x^{-1} and y^{-1} . Let $\cdot : F_2 \times F_2 \rightarrow F_2$ denote the concatenation–reduction operation. Then $(F_2, \cdot)$ is a group: the **free group** of rank 2.*

This group, which we've renamed F_2 and given a special name to, is the most general possible group we can make from two generators x and y . Before we go any further, we state some formal definitions.

Definition 5.3 Let $X = \{x_1, x_2, \dots\}$ be a (possibly infinite) set of formal symbols, and let $X^{-1} = \{x_1^{-1}, x_2^{-1}, \dots\}$ be a set of corresponding formal inverses.

Denote by $X^\pm = X \cup X^{-1}$ the union of these two sets; we call this an **alphabet**.

A **word** in $X^\pm$ is a (possibly empty) string of symbols from $X^\pm$; the **empty word** is denoted by e .

A word is said to be **reduced** if all possible cancellations of adjacent symbols and their inverses have been performed; we may choose to write a reduced word in an even more compact form by using exponents to group adjacent symbols of the same type.

Denote the set of reduced words in $X^\pm$ by $W(X^\pm)$.

This set $W(X^\pm)$ can be partitioned into a number of pairwise disjoint subsets in a fairly natural way. Since the reduced form of a given word is unique, it makes sense to talk about its **length**: the number of symbols in the string we get if we ungroup all of the adjacent repetitions (but don't introduce any empty words or cancellable pairs). For a given word $w \in W(X^\pm)$ we denote its length by $l(w)$.

For example, $l(x^3yx^{-1}) = l(xxxyx^{-1}) = 5$, since when we write the word out in full we end up with a string of five symbols. We define the length of the empty word to be zero: $l(e) = 0$.

Using this concept we can partition $W(X^\pm)$ into the following subsets:

$$\begin{aligned} W_0(X^\pm) &= \{w \in W(X^\pm) : l(w) = 0\} = \{e\} \\ W_1(X^\pm) &= \{w \in W(X^\pm) : l(w) = 1\} = X^\pm \\ W_2(X^\pm) &= \{w \in W(X^\pm) : l(w) = 2\} \\ &\vdots \\ W_n(X^\pm) &= \{w \in W(X^\pm) : l(w) = n\} \end{aligned}$$

We're now ready to define the free group $F(X)$ on a set X of symbols. There are two different ways of doing this, one concrete and one abstract. The concrete, constructive definition is essentially a formalisation and generalisation of the preceding discussion.

Definition 5.4 Let $X = \{x_1, x_2, \dots\}$ be a (possibly infinite) set of symbols. Then the **free group** $F(X)$ with **basis** or **generating set** X is the set $W(X^\pm)$ of reduced words in the symbols $x_1, x_2, \dots$ and their formal inverses $x_1^{-1}, x_2^{-1}, \dots$, equipped with the binary operation $\cdot : W(X^\pm) \times W(X^\pm) \rightarrow W(X^\pm)$ given by the concatenation–reduction operation described above.

The **rank** of $F(X)$ is the cardinality of the set X ; in the finite case $|X| = n$ we will often denote the free group $F(X)$ by F_n .

The other definition is more abstract, and it's not immediately obvious that it's equivalent (or even that the objects it describes actually exist). Nevertheless, there are certain technical advantages to this definition as well, and some books present this as the main definition and relegate the previous one to a concrete construction.

Definition 5.5 Let X be a set. The **free group** with **basis** X is the group satisfying the property that for any other group G and any function $f: X \rightarrow G$, there is a unique homomorphism $\tilde{f}: F(X) \rightarrow G$ which **extends** f . That is, $\tilde{f}(x) = f(x) \in G$ for any $x \in X$. Equivalently, the diagram

$$\begin{array}{ccc} X & \xrightarrow{f} & G \\ \downarrow i & \nearrow \tilde{f} & \\ F(X) & & \end{array}$$

commutes, in the sense that $\tilde{f} \circ i(x) = f(x)$ for all $x \in X$. (Here $i: X \rightarrow F(X)$ is the obvious inclusion map.)

This is a generalisation of a situation you might be familiar with from linear algebra. Suppose $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots\}$ is a set of vectors in some vector space V . Then for any other vector space W and function $f: S \rightarrow W$ there's a unique linear map $\tilde{f}: \text{span}(S) \rightarrow W$ which extends f to the span of S (that is, the subspace of V consisting of all possible linear combinations of the vectors in S). This linear map $\tilde{f}$ is determined completely by what f does to each of the vectors in S , and much the same thing happens in the corresponding scenario for groups. In a sense, the free group $F(X)$ on a set X is a sort of group-theoretic analogue of the linear algebraic concept of the span of a set of vectors.

Definition 5.5 contains some important subtlety. The functions $f: X \rightarrow G$ and $i: X \hookrightarrow F(X)$ map an object of one type (the set X) to an object of another type (the groups G and $F(X)$). Strictly speaking, this isn't really allowed: functions map sets to sets, and homomorphisms map groups to groups.

Happily, we can resolve this apparent type mismatch error and gain a deeper understanding of how the free group construction works. The key is to notice that the functions $f: X \rightarrow G$ treat the group G just as an ordinary set, ignoring any extra structure it might have. The group structure only becomes relevant when we look at the extended homomorphism $\tilde{f}: F(X) \rightarrow G$, so to be really precise we need a way of clarifying exactly when we're talking about the group G and when we're talking about its underlying set.

Let $U(G)$ be the underlying set of G ,⁴ without any group structure. Then Definition 5.5 says that functions $f: X \rightarrow U(G)$ are in bijective correspondence with homomorphisms $\tilde{f}: F(X) \rightarrow G$.

If we denote by $\text{Hom}_{\text{Set}}(X, Y)$ the set of all possible functions from the set X to the set Y , and by $\text{Hom}_{\text{Group}}(G, H)$ the set of all possible homomorphisms from the group G to the group H , then we can write this as

$$\text{Hom}_{\text{Set}}(X, U(G)) \cong \text{Hom}_{\text{Group}}(F(X), G).^6$$

We'll make use of this fact a little later, but now we can show that Definitions 5.4 and 5.5 are equivalent.

Proposition 5.6 *Let $X = \{x_1, x_2, \dots\}$ be a (possibly infinite) set, and let $F(X)$ be the free group of reduced words in the alphabet $X^\pm$. For any group G and function $f: X \rightarrow G$ there is a unique homomorphism $\tilde{f}: F(X) \rightarrow G$ which extends f .*

Conversely, let F be a group and let $X \subset F$ be some subset of F . If, for any group G and function $f: X \rightarrow G$ there exists a unique homomorphism $\tilde{f}: F \rightarrow G$ which extends f , then F is isomorphic to the free group $F(X)$ of reduced words in the alphabet $X^\pm$.

Proof Every element w of $F(X)$ has a unique expression as a reduced word in the elements of the alphabet $X^\pm$:

$$w = x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k}$$

where $n_1, \dots, n_k \in \mathbb{Z}$ and $x_{i_1}, \dots, x_{i_k} \in X$ are a collection of elements of X (some of which might be the same as each other).

Then, given a function $f: X \rightarrow G$, the only way we can extend this to a homomorphism $\tilde{f}: F(X) \rightarrow G$ is by mapping the word

$$w = x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k}$$

to the product

$$f(x_{i_1})^{n_1} * f(x_{i_2})^{n_2} * \dots * f(x_{i_k})^{n_k}$$

in G , and mapping the empty word e in $F(X)$ to the identity e_G in G . Nothing else satisfies the homomorphism condition. So, the required homomorphism $\tilde{f}$ exists and is unique.

To prove the converse, suppose that for any $f: X \rightarrow G$ there exists a unique homomorphism $\tilde{f}: F \rightarrow G$ extending f . Let $\phi: F(X) \rightarrow F$ be the unique homomorphism which maps any reduced word in $F(X)$ to the element of F obtained by regarding this word as a product of the elements of F . That is, suppose $(F, \diamond)$ is the hypothesised group. Then for any $w = x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k}$ in $F(X)$ we set

$$\phi(x_{i_1}^{n_1} x_{i_2}^{n_2} \dots x_{i_k}^{n_k}) = x_{i_1}^{n_1} \diamond x_{i_2}^{n_2} \diamond \dots \diamond x_{i_k}^{n_k}.$$

⁴ We can regard U as a device that turns groups into sets, and we can view the free group construction F as a machine that turns sets into groups. More than that, U turns group homomorphisms into functions, and F turns functions into group homomorphisms. So U and F translate neatly between the world of sets and the world of groups.

Maps of this type are called **functors**. The free group construction $F: \text{Set} \rightarrow \text{Group}$ is a functor from the **category** of sets to the category of groups. The underlying set map $U: \text{Group} \rightarrow \text{Set}$ is a functor from the category of groups to the category of sets; it is an example of something called a **forgetful functor**.

A full discussion of categories and functors is beyond the scope of this book, but the branch of mathematics in which they play a fundamental part, called **category theory** (but sometimes flip-pantly referred to as “abstract nonsense”) is at the same time perhaps the most abstract subject in modern mathematics and one of the most powerful: it provides a suitably precise language and framework within which to discuss many different sub-branches of mathematics in generality.

Much of algebraic topology, for example, one of the highlights of twentieth- and twenty-first century mathematics, is concerned with the study of functors from categories of topological objects to categories of algebraic objects such as groups and rings. Category theory also has important applications in theoretical computer science.

Interested readers are directed to the classic work *Categories for the Working Mathematician* by Saunders Mac Lane (1909–2005),⁵ one of the main pioneers of the field. Many years after its original publication in 1971, it remains one of the most comprehensive and lucid treatments of the subject.

⁵ S Mac Lane, *Categories for the Working Mathematician*, second edition, Graduate Texts in Mathematics 5, Springer (1998)

⁶ This is an example of something category theorists call an **adjunction** or **adjoint relationship**. The functor F is said to be a **left adjoint** of U , and U is a **right adjoint** of F .

We want to show that ϕ is an isomorphism. To do this, let $i: X \hookrightarrow F(X)$ be the obvious inclusion function. Then we can extend this to a unique homomorphism $\psi = \tilde{i}: F \rightarrow F(X)$.

The composition $\psi \circ \phi: F(X) \rightarrow F(X)$ is the identity homomorphism on any element x_i of X , since ϕ maps x_i in X to the corresponding element in F , and ψ maps it straight back to the corresponding element of $X \subset F(X)$, since ψ is just the extension of the inclusion map $i: X \hookrightarrow F(X)$ to the whole of F .

Since $\psi \circ \phi$ is the identity on X it must also be the identity on $F(X)$ since if all the elements of X are mapped to themselves, then so must all possible reduced words made from them and their inverses.

Also, $\phi \circ \psi$ must be the identity on F . Consider the inclusion map $j: X \hookrightarrow F$. By the hypothesis, this extends to the identity homomorphism $\text{id}_F: F \rightarrow F$. But $\phi \circ \psi$ also extends j to all of F . The hypothesis said that any such function extends to a unique homomorphism, so the composite $\phi \circ \psi$ must equal the identity map id_F .

We've now shown that there exist homomorphisms $\phi: F(X) \rightarrow F$ and $\psi: F \rightarrow F(X)$, with $\phi \circ \psi = \text{id}_F$ and $\psi \circ \phi = \text{id}_{F(X)}$. This is equivalent to saying that both ϕ and ψ are isomorphisms, and so $F(X) \cong F$. $\square$

In Definitions 5.4 and 5.5 we referred to *the* free group with a particular basis, a phrasing that quietly implies that there is just one such group with that basis. The following proposition justifies this:

Proposition 5.7 *Suppose that X_1 and X_2 are (possibly infinite) sets such that there exists a bijection $f: X_1 \rightarrow X_2$. Then $F(X_1) \cong F(X_2)$.*

Conversely, suppose that $F(X_1)$ and $F(X_2)$ are the free groups generated by two (possibly infinite) sets X_1 and X_2 , and that $F(X_1) \cong F(X_2)$. Then there exists a bijection $f: X_1 \rightarrow X_2$.

Proof Let $i_1: X_1 \hookrightarrow F(X_1)$ and $i_2: X_2 \hookrightarrow F(X_2)$ be the appropriate inclusion maps.

We can compose i_1 with f^{-1} to get a function $i_1 \circ f^{-1}: X_2 \rightarrow F(X_1)$, and we can compose i_2 with f to get a function $i_2 \circ f: X_1 \rightarrow F(X_2)$. By the fundamental mapping property of free groups, we can extend $i_1 \circ f^{-1}$ to get a unique homomorphism $\psi: F(X_2) \rightarrow F(X_1)$, and we can also extend $i_2 \circ f$ to get a unique homomorphism $\phi: F(X_1) \rightarrow F(X_2)$, as depicted by the following diagrams:

$$\begin{array}{ccc}
 X_1 & \xrightarrow{f} & X_2 \hookrightarrow F(X_2) \\
 i_1 \downarrow & \nearrow \phi & \\
 F(X_1) & &
 \end{array}
 \qquad
 \begin{array}{ccc}
 X_2 & \xrightarrow{f^{-1}} & X_1 \hookrightarrow F(X_1) \\
 i_2 \downarrow & \nearrow \psi & \\
 F(X_2) & &
 \end{array}$$

(Compare these diagrams with the one in Definition 5.5.)

For any $x \in X_1$ we have $\psi \circ \phi(x) = \psi(f(x)) = f^{-1}(f(x)) = x$. So $\psi \circ \phi$ extends the inclusion map $i_1: X_1 \hookrightarrow F(X_1)$ to the whole of $F(X_1)$. But so does the identity map $\text{id}_{F(X_1)}: F(X_1) \rightarrow F(X_1)$, hence $\psi \circ \phi = \text{id}_{F(X_1)}$. A similar argument shows that $\phi \circ \psi = \text{id}_{F(X_2)}$ and hence both ϕ and ψ are the required isomorphisms.

To prove the converse, suppose that $F(X_1) \cong F(X_2)$. Then for any group G , the set $\text{Hom}_{\text{Group}}(F(X_1), G)$ of homomorphisms $F(X_1) \rightarrow G$ is in bijective correspondence with the set $\text{Hom}_{\text{Group}}(F(X_2), G)$ of group homomorphisms $F(X_2) \rightarrow G$.

By the fundamental mapping property of free groups, we also know that since $\text{Hom}_{\text{Group}}(F(X_1), G)$ is in bijective correspondence with the set $\text{Hom}_{\text{Set}}(X_1, U(G))$ of functions $X_1 \rightarrow G$ and $\text{Hom}_{\text{Group}}(F(X_2), G)$ is in bijective correspondence with the set $\text{Hom}_{\text{Set}}(X_2, U(G))$ of functions $X_2 \rightarrow G$, it must follow that the sets $\text{Hom}_{\text{Set}}(X_1, U(G))$ and $\text{Hom}_{\text{Set}}(X_2, U(G))$ are in bijective correspondence with each other.

In particular, this means that they must have the same cardinality:

$$|\text{Hom}_{\text{Set}}(X_1, U(G))| = |\text{Hom}_{\text{Set}}(X_2, U(G))|$$

In general, the number of possible functions from a set A with cardinality a to a set B with cardinality b is given by b^a . This is because each of the elements of A has b possible elements in B that it could map to, so there are $b \times \cdots \times b = b^a$ possible distinct functions $A \rightarrow B$. Bearing this in mind, if we choose G to be some finite group, say $\mathbb{Z}_2$, then we can see that

$$|\text{Hom}_{\text{Set}}(X_1, U(\mathbb{Z}_2))| = |\mathbb{Z}_2|^{|X_1|} = 2^{|X_1|}$$

and

$$|\text{Hom}_{\text{Set}}(X_2, U(\mathbb{Z}_2))| = |\mathbb{Z}_2|^{|X_2|} = 2^{|X_2|}$$

so $2^{|X_1|} = 2^{|X_2|}$ and therefore $|X_1| = |X_2|$ as required. $\square$

The final step in the second half of this proof works for infinite sets as well as finite ones, although a detailed discussion would be a bit too much of a digression here. For finite sets it's obvious: just take base-2 logarithms of both sides. It would also have worked if we'd replaced $\mathbb{Z}_2$ with $\mathbb{Z}_3$, V_4 , a dihedral group D_n , or any other nontrivial finite group.

Corollary 5.8 *Let $m, n \in \mathbb{N}$ be two positive integers. Then $F_m \cong F_n$ if and only if $m = n$.*

The really important thing this proposition (and the immediate corollary which follows it) tells us is that the **rank** of a free group (the cardinality of its basis) is well-defined: any other set with the same cardinality will generate an isomorphic free group. This is analogous to the fact that the dimension of a vector space is well-defined: given

any finite-dimensional vector space V over some field $\mathbb{K}$, any two bases for V will have the same number of elements.

Next we'll look at the order $|w|$ of a given word w in a free group, but first we need to introduce a new concept, a slight variation on the idea of a reduced word:

Definition 5.9 Let $w = x_{i_1} \dots x_{i_k}$ be a reduced word in a free group $F(X)$ on some set X , where $x_{i_1}, \dots, x_{i_k} \in X^\pm$. Then w is **cyclically reduced** if $x_{i_1} \neq x_{i_k}^{-1}$. That is, if we cyclically permute w to get a new word $v = x_{i_k} x_{i_1} \dots x_{i_{k-1}}$ then this new word v is also reduced.

For example, the word

$$x_1^3 x_2^{-1} x_1 x_3^2 x_2^{-1} = x_1 x_1 x_1 x_2^{-1} x_1 x_3 x_3 x_2^{-1}$$

in F_3 is cyclically reduced, but the word

$$x_1^3 x_2^{-1} x_1 x_3^2 x_2^{-1} x_1^{-2} = x_1 x_1 x_1 x_2^{-1} x_1 x_3 x_3 x_2^{-1} x_1^{-1} x_1^{-1}$$

isn't, since a single cyclic permutation yields $x_1^{-1} x_1^3 x_2^{-1} x_1 x_3^2 x_2^{-1} x_1^{-1}$ which reduces to $x_1^2 x_2^{-1} x_1 x_3^2 x_2^{-1} x_1^{-1}$, which in turn is reduced but not cyclically reduced.

Recall that an element $g \neq e$ of some group G is a **torsion element** if its order $|g|$ is finite.⁷ A group G with no elements of finite order (except for the identity e) is said to be **torsion free**. Free groups have no torsion elements:

Proposition 5.10 Let F be a free group generated by some set X . Then F is torsion free.

Proof Let $w \in F$ be a nontrivial reduced word in the alphabet $X^\pm$. Any such word is conjugate to a cyclically reduced word, in the sense that $w = uvu^{-1}$ for some $u, v \in F$ where v is cyclically reduced.

If w is already cyclically reduced then $u = e$ is the empty word, otherwise u consists of the subword at the beginning of w that cancels with a corresponding inverse subword at the end of w , while v is the non-cancelling middle section of w .

If w is cyclically reduced, then so is w^2 , and for that matter w^n for any $n \in \mathbb{N}$. So the length $l(w^n)$ of n concatenated copies of a cyclically reduced word w is equal to $n \cdot l(w)$.

Now we consider the length of w and its subwords u and v . Let $r = l(w)$ and $s = l(u)$. Then $l(v) = r - 2s$, which is positive if $w \neq e$. Then $w^n = (uvu^{-1})^n = uv^n u^{-1}$ and hence

$$l(w^n) = nl(v) + 2l(u) = n(r - 2s) + 2s.$$

This can't be less than n , so $l(w^n) \geq n$ for all $n \geq 1$.

In particular, this means that for a nonempty word w , any positive power w^n can't have zero length, which means that it can't be the

⁷ Definition 1.18, page 11.

empty word. Hence there is no $n \in \mathbb{N}$ for which $w^n = e$ and so w can't have finite order. Thus F has no torsion elements. $\square$

The following important theorem is introduced now, but the proof is rather technical and we'll postpone the details until Section 5.B.

Theorem 5.11 (The Nielsen–Schreier Theorem) *Let F be a free group, and let H be a subgroup of F . Then H is itself a free group. Furthermore, if the index $|F : H| = s$ of H in F and $\text{rank}(F) = r$ are both finite, then*

$$\text{rank}(H) = (r - 1)s + 1.$$

The next proposition tells us that except for the case of the infinite cyclic group $\mathbb{Z} \cong F_1$, free groups are as nonabelian as they can be.

Proposition 5.12 *Let F be a free group generated by some set X , and consider two elements $u, v \in F$. Then $uv = vu$ if and only if there exists some $w \in F$ and $m, n \in \mathbb{Z}$ such that $u = w^m$ and $v = w^n$.*

Proof The “if” direction is obvious: suppose $u = w^m$ and $v = w^n$. Then

$$uv = w^m w^n = w^{m+n} = w^n w^m = vu.$$

The “only if” direction is more involved. Let u and v be reduced words in F , and let $r = l(u)$ and $s = l(v)$. We proceed by induction on $r + s$, and without loss of generality we can assume that $0 \leq r \leq s$.

If $r = 0$ then $u = e = v^0$ so we can set $w = v$ and let $m = 0$ and $n = 1$. This confirms the induction basis.

Now suppose that $r > 0$. If we concatenate u and v then it may be the case that part of the end of u cancels with part of the beginning of v . Let h be the length of this cancelling subword at the end of u , and equivalently the length of the cancelling subword at the beginning of v . (It's entirely possible that $h = 0$ if no cancellation occurs.)

Then $l(uv) = l(u) + l(v) - 2h = r + s - 2h$. Since $uv = vu$ we must also have $2h$ letters cancelling when we concatenate v and u . We have three cases to consider:

Case 1 ($h = 0$) Here there are no cancellations when we concatenate u and v or when we concatenate v and u . So

$$uv = u_1 \dots u_r v_1 \dots v_r \dots v_s = v_1 \dots v_r \dots v_s u_1 \dots u_r = vu$$

tells us that $u_1 = v_1, u_2 = v_2$, all the way up to $u_r = v_r$. This means we can write $v = uz$ where $z = v_{r+1} \dots v_s$. Since $uv = vu$ it follows that v must be in the centraliser $Z_F(u)$ of the element u in F . But $u \in Z_F(u)$ as well, so $z = u^{-1}v$ must also be in $Z_F(u)$ and hence $uz = zu$.

The length $l(z) = s - r < l(v)$ and so by the induction hypothesis applied to u and z it follows that $u = w^p$ and $z = w^q$ for some $w \in F$ and $p, q \in \mathbb{Z}$. Then $u = w^p$ and $v = w^{p+q}$ as claimed.

Case 2 ($h = r$) Here, the cancelling portion of u is the entirety of u , which means that $v = u^{-1}z$ for some word $z \in F$, where $l(z) = s - r < l(v)$. We can apply a similar argument to the one we used in the previous case, by appealing once more to the inductive hypothesis. We have $uv = vu$ which means that $uu^{-1}z = u^{-1}zu$, and hence $z = u^{-1}zu$, so $uz = zu$. Therefore $u = w^p$ and $z = w^q$ for some $w \in F$ and $p, q \in \mathbb{Z}$, so $u = w^p$ and $v = u^{-1}z = w^{q-p}$.

Case 3 ($0 < h < r$) Here we have

$$uv = u_1 \dots u_{r-h} v_{h+1} \dots v_s$$

$$\text{and} \quad vu = v_1 \dots v_{s-h} u_{h+1} \dots u_r.$$

By the hypothesis, these are equal, and so in particular we have $u_1 = v_1$ and $u_r = v_s$. Since $h > 0$ we have at least some cancellation, so that tells us that $u_r = v_1^{-1}$ and $v_s = u_1^{-1}$. Putting this together we see that

$$u_r = v_1^{-1} = u_1^{-1} = v_r$$

and thus there exist words $a, b \in F$ such that $u = u_1 a u_1^{-1}$ and $v = u_1 b u_1^{-1}$, where

$$a = u_2 \dots u_{r-1}, \quad b = v_2 \dots v_{s-1},$$

$$l(a) = r - 2, \quad l(b) = s - 2.$$

Then $uv = vu$ means that $u_1 a b u_1^{-1} = u_1 b a u_1^{-1}$, which implies that $ab = ba$. By the induction hypothesis we then have $a = z^p$ and $b = z^q$ for some $p, q \in \mathbb{Z}$ and $z \in F$. Hence

$$u = u_1 z^p u_1^{-1} = (u_1 z u_1^{-1})^p,$$

$$v = u_1 z^q u_1^{-1} = (u_1 z u_1^{-1})^q.$$

Setting $w = (u_1 z u_1^{-1})$, $m = p$ and $n = q$ the result follows. $\square$

What this proposition tells us is that no two words u and v in a free group F commute with each other unless they happen to both be powers of the same element. A free group F of rank 2 or greater is therefore no more abelian than it absolutely has to be.

For the relations of words are in pairs first.

For the relations of words are sometimes in oppositions.

For the relations of words are according to their distances from the pair.

— Christopher Smart (1722–1771),
Jubilate Agno (1759–1763)
 Fragment B, part 4

5.2 Generators and relations

IN THIS SECTION we will develop the machinery of group presentations. Our first ingredient is the next proposition, which shows that any group can be regarded as a quotient of a free group.

Proposition 5.13 *Let G be a group, and let X be a set of generators for G . Then G is isomorphic to a quotient of the free group $F(X)$.*

Proof There is a well-defined homomorphism $f: F(X) \rightarrow G$ which maps any word $w = x_{i_1}^{n_1} \dots x_{i_k}^{n_k}$ in $F(X)$ onto the corresponding product of generators in G .

This homomorphism is surjective, since G is generated by the elements of X . Then

$$F(X)/\ker(f) \cong \text{im}(f) = G$$

by the First Isomorphism Theorem. $\square$

We now have all we need to formally define the notion of a group presentation: we have a concrete definition of the **free group** on a set of generators, and Proposition 5.13 tells us that any group is isomorphic to a quotient of such a group.

Going back to our motivating example of the finite cyclic group $\mathbb{Z}_n = \{1, t, t^2, \dots, t^{n-1}\}$ (which for the moment we'll write using multiplicative notation), this is constructed from a single generator t subject to the relation $t^n = 1$. In the context of Proposition 5.13 we can think of this as the free group $F_1 \cong \mathbb{Z}$ on a single generator t , modulo some normal subgroup which we can think of as the kernel of some surjective homomorphism $F_1 \rightarrow \mathbb{Z}_n$.

By the fundamental property of free groups, any such homomorphism is completely determined by what its generators map to. In this case there's only one generator t to worry about, and as long as we map it to a generator of $\mathbb{Z}_n$, the corresponding homomorphism will be surjective.

So let's take the homomorphism $f: F_1 \rightarrow \mathbb{Z}_n$ which maps the generator $t \in F_1$ to the generator $t \in \mathbb{Z}_n$. What is the kernel of this homomorphism? It's everything in F_1 that maps to the identity $1 \in \mathbb{Z}_n$. But these elements are exactly the powers t^k of t for which k is a multiple of n .

By Proposition 5.13 we know that $\mathbb{Z}_n \cong F_1/\ker(f)$ where $\ker(f) = \{t^k : n|k\} \triangleleft F_1$.

This says that the kernel $\ker(f)$ consists of all the words in F_1 which are powers of the word t^n , in other words $\ker(f)$ is generated by the word t^n .

We could just as easily write this scenario as

$$\mathbb{Z}_n \cong \langle t \rangle / \langle t^n \rangle$$

but we'll usually adopt the notation we saw briefly at the beginning of the chapter,

$$\mathbb{Z}_n \cong \langle t : t^n \rangle$$

or, equivalently,

$$\mathbb{Z}_n \cong \langle t : t^n \rangle$$

when it's understood implicitly that the word or words to the right of the colon are equal to the identity.

The group $\mathbb{Z} \oplus \mathbb{Z}$ can be constructed by taking a suitable free group, in this case $F_2 = \langle x, y \rangle$ and introducing some appropriate relations that force everything to commute. This is in some senses fairly straightforward, because all we need to do is to ensure that the generators x and y commute, and since every other element of our desired group can be regarded as a reduced word in these generators and their inverses, all such elements will commute.

This can be effected by imposing the relation $xy = yx$, to get the presentation

$$\mathbb{Z} \oplus \mathbb{Z} \cong \langle x, y : xy = yx \rangle.$$

Alternatively, replacing the relation $xy = yx$ with the relator $xyx^{-1}y^{-1}$ (which is also the commutator $[x, y]$) gives us

$$\mathbb{Z} \oplus \mathbb{Z} \cong \langle x, y : xyx^{-1}y^{-1} \rangle.$$

Another way of looking at this is to observe that $xyx^{-1}y^{-1}$ is a generator for the commutator subgroup $[F_2, F_2]$, which is exactly the kernel of the projection homomorphism $F_2 \twoheadrightarrow \mathbb{Z} \oplus \mathbb{Z} = F_2/[F_2, F_2]$. We'll come back to this in the next section, when we study finitely-generated abelian groups.

But now, before looking at any more examples, it's time to introduce some formal definitions.

Definition 5.14 Let G be a group and let S be some subset of G . The **normal closure** $\bar{S}$ or $\langle S^G \rangle$ of S in G is the intersection of all normal subgroups of G which contain S .

The following proposition gives a constructive definition for $\langle S^G \rangle$:

Proposition 5.15 Let G be a group and let S be some subset of G . Then the normal closure

$$\langle S^G \rangle = \{g * s * g^{-1} : g \in G \text{ and } s \in S\}.$$

Proof By Proposition 3.10 we know that any normal subgroup is closed under conjugation. So any normal subgroup of G which contains S must also contain all of its conjugates, hence $\{g * s * g^{-1} : g \in G \text{ and } s \in S\} \subseteq \langle S^G \rangle$.

But the set $\{g * s * g^{-1} : g \in G \text{ and } s \in S\}$ is itself a normal subgroup of G containing S : set $g = e$, then $g * s * g^{-1} = e * s * e^{-1} = s$ for all $s \in S$. Hence $\langle S^G \rangle \subseteq \{g * s * g^{-1} : g \in G \text{ and } s \in S\}$ and so the two are equal. $\square$

We can now define the concept of a group presentation:

Definition 5.16 Let $X = \{x_1, x_2, \dots\}$ be a (finite, infinite or empty) set of generators, and let $R = \{w_1, w_2, \dots\}$ be a (finite, infinite or empty) set of reduced words in the free group $F(X)$. Suppose $G \cong F(X)/\langle R^{F(X)} \rangle$.

Then the elements of X are called **generators** for G , the words in R are called **relators** for G , and we write the corresponding **presentation** (or **free presentation**) of G as

$$G \cong \langle X : R \rangle = \langle x_1, x_2, \dots : w_1, w_2, \dots \rangle.$$

A **relation** in G is an equation of the form (or equivalent to) $w = e$, where $w \in R$ and e is the identity element of G . We can also write a presentation for G by using relations instead of relators:

$$G \cong \langle x_1, x_2, \dots : w_1 = w_2 = \dots = e \rangle$$

If R is empty, then the presentation $\langle X : R \rangle = \langle X : \emptyset \rangle = \langle X : \rangle$ gives the free group $F(X)$. If both X and R are finite, then we say that the group $G = \langle X : R \rangle$ is **finitely presented**.

Some books use a vertical bar instead of a colon as the delimiter:

$$G \cong \langle X \mid R \rangle$$

Group presentations satisfy a similar universal mapping property to the one satisfied by free groups:⁸

⁸ Proposition 5.6, page 139.

Proposition 5.17 Let $G = \langle X : R \rangle$ be the group determined by the given presentation, and let H be any group. Suppose that $f: X \rightarrow H$ is a function satisfying the property that $f(x^{-1}) = f(x)^{-1}$ for any $x \in X$, and that for any relator $w = a_1 \dots a_k$ in R , where $a_1, \dots, a_k \in X^\pm$, we have $f(a_1) \dots f(a_k) = e_H \in H$. Then the function f extends uniquely to a homomorphism $\hat{f}: G \rightarrow H$.

Proof If the function $f: X \rightarrow H$ extends to a group homomorphism $\hat{f}: G \rightarrow H$ at all, then it must be the case that

$$f(a_1 \dots a_k) = f(a_1) \dots f(a_k)$$

for all $a_1, \dots, a_k \in X^\pm$, which means that $\hat{f}$ must be unique.

Now we have to show that $\hat{f}$ actually exists. Proposition 5.6 says that $f: X \rightarrow H$ extends to a unique homomorphism $\tilde{f}: F(X) \rightarrow H$ which extends f in the sense that $\tilde{f}|_X = f$.

The relator set R is contained in the kernel of $\tilde{f}$ since, by hypothesis,

$$f(a_1 \dots a_k) = f(a_1) \dots f(a_k) = e_H \in H$$

for any word $a_1 \dots a_k \in R$. Let $N = \langle R^{F(X)} \rangle$ be the normal closure of R in $F(X)$. This is a subgroup of $\ker(\tilde{f})$, which is in turn a normal subgroup of the $F(X)$, and by definition $G = F(X)/N$.

Now define $\hat{f}: G = F(X)/N \rightarrow H$ by $\hat{f}((a_1 \dots a_k)N) = \tilde{f}(a_1 \dots a_k)$. Then $\hat{f}$ is a homomorphism, and $\hat{f}(xN) = \tilde{f}(x) = f(x)$ for all $x \in X$, so it extends f as required. $\square$

We can write down a presentation for $\mathbb{Z} \oplus \mathbb{Z}$ from the generators x and y together with the relator $xyx^{-1}y^{-1}$:

$$\langle x, y : xyx^{-1}y^{-1} \rangle$$

or the relation $xyx^{-1}y^{-1}$:

$$\langle x, y : xyx^{-1}y^{-1} = e \rangle$$

or the equivalent relation $xy = yx$:

$$\langle x, y : xy = yx \rangle$$

This is all well and good, and we've already seen that the language of presentations, generators and relations is a useful and compact way of describing at least some groups, but for it to be really useful we need to know what groups can be described in this way. The answer, happily, is that all of them can:

Proposition 5.18 *Every group G has a presentation $\langle X : R \rangle$ where X is a (finite, infinite or empty) set of generators, and R is a (finite, infinite or empty) set of relators. Furthermore, if G is finite, then there exists a finite presentation for G .*

Proof Let $X \subset G$ be a set of generators for G . (Setting $X = G$ will work, for example, but there will typically be a smaller set that suffices.) Then by Proposition 5.13 there exists a homomorphism $f: F(X) \rightarrow G$ such that $G \cong F(X) / \ker(f)$, and by setting R to be a set of generators for $\ker(f)$ we obtain a presentation

$$\langle X : R \rangle \cong G.$$

(Setting $R = \ker(f)$ itself will work, but there will usually be a smaller set which generates $\ker(f)$.)

If G is finite, with $|G| = s$ for some $s \in \mathbb{N}$, then there exists a finite generating set X , with $|X| = r$ for some $r \in \mathbb{N}$.

By the Nielsen–Schreier Theorem⁹ there exists a set R which generates $\ker(f)$, with $|R| = (r - 1)s + 1$. Hence

$$G \cong \langle X : R \rangle$$

is a finite presentation for G . □

Thus every (finite or infinite) group can be expressed as a presentation, and while the above proposition merely assures the existence of such a presentation, it's a relatively straightforward process to write one down, as the following example shows.

Example 5.19 Let $V_4 = \{e, a, b, c\}$ be the Klein group as usual. We can write down a presentation for this group by using V_4 itself as the set of generators, and reading off the relations from the multiplication table (Table 5.1).

	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

Table 5.1: Multiplication table for the Klein group V_4

⁹Theorem 5.11, page 143.

This gives us the presentation

$$V_4 \cong \left\langle e, a, b, c : \begin{array}{llll} ee = e & ea = a & eb = b & ec = c \\ ae = a & aa = e & ab = c & ac = b \\ be = b & ba = c & bb = e & bc = a \\ ce = c & ca = b & cb = a & cc = e \end{array} \right\rangle.$$

The main drawback with this method is that although it does provide a presentation for a given group, that presentation will usually not be the simplest one. And for an infinite group, this method becomes completely impractical for obvious reasons. There will almost always be a simpler presentation for the groups we're likely to meet. For example, the above presentation for V_4 has four generators and sixteen relations, but we really only need two generators and three relations:

Example 5.20 The Klein group V_4 is isomorphic to the direct sum $\mathbb{Z}_2 \oplus \mathbb{Z}_2$. We already know how to write down a presentation for

$$\mathbb{Z} \oplus \mathbb{Z} \cong \langle x, y : xy = yx \rangle,$$

and we can use this to make a presentation for $\mathbb{Z}_2 \oplus \mathbb{Z}_2$. We do this by introducing two more relations $x^2 = e$ and $y^2 = e$, thus forcing each generator to have order 2. This gives us the presentation

$$V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \cong \langle x, y : x^2 = y^2 = e, xy = yx \rangle.$$

In general we will want a suitably compact presentation of a given group, or at least one that can be written down in a compact way.

Example 5.21 Let $G = \mathbb{Z}_3 \oplus \mathbb{Z}_2$. Then

$$G \cong \langle x, y : x^3 = y^2 = e, xy = yx \rangle.$$

Examining the presentations in these two examples, we see that each group is a direct sum of finite cyclic groups, and each has a presentation comprising the combined generators and relations for the cyclic group summands together with another relation that forces the generator of one to commute with the generator of the other. That is,

$$\begin{aligned} \langle x : x^2 = e \rangle \oplus \langle y : y^2 = e \rangle &= \langle x, y : x^2 = e, y^2 = e, xy = yx \rangle, \\ \langle x : x^3 = e \rangle \oplus \langle y : y^2 = e \rangle &= \langle x, y : x^3 = e, y^2 = e, xy = yx \rangle. \end{aligned}$$

This illustrates the following proposition:

Proposition 5.22 Suppose that $\langle X : R \rangle$ is a presentation for a group G , and that $\langle Y : S \rangle$ is a presentation for a group H , with $X \cap Y = \emptyset$. Then

$$\langle X \cup Y : R \cup S \cup [X, Y] \rangle$$

is a presentation for $G \times H$ (or $G \oplus H$), where

$$[X, Y] = \{xyx^{-1}y^{-1} : x \in X, y \in Y\}$$

is the set of commutators of elements from X and Y .

Proof Suppose

$$K = \langle X \cup Y : R \cup S \cup [X, Y] \rangle.$$

Define a function $f: X \cup Y \rightarrow G \times H$ by $f(x) = (x, e_H)$ and $f(y) = (e_G, y)$ for all $x \in X$ and $y \in Y$.

Let $w = a_1 \dots a_k \in R \cup S \cup [X, Y]$ be a relator. We need to check that $f(w) = (e_G, e_H)$, and there are three cases to consider.

If $w \in R$ then $w \in F(X)$ and so $f(w) = (w, e_H) = (e_G, e_H)$ since $w = e_G$ in G .

Similarly, if $w \in S$ then $w \in F(Y)$ and so $f(w) = (e_G, w) = (e_G, e_H)$ since $w = e_H$ in H .

If, on the other hand, $w \in [X, Y]$ then $w = xyx^{-1}y^{-1}$ for some $x \in X$ and $y \in Y$. Hence

$$\begin{aligned} f(w) &= f(xyx^{-1}y^{-1}) = f(x)f(y)f(x^{-1})f(y^{-1}) \\ &= (x, e_H)(e_G, y)(x^{-1}, e_H)(e_G, y^{-1}) = (e_G, e_H). \end{aligned}$$

So $f(w) = e_{G \times H}$ for any relator $w \in R \cup S \cup [X, Y]$, and hence by the fundamental mapping property of group presentations,¹⁰ f extends uniquely to a homomorphism $\hat{f}: K \rightarrow G \times H$.

The function $f: X \cup Y \rightarrow G \times H$ is surjective, and hence so is the homomorphism $\hat{f}: K \rightarrow G \times H$.

Since $xy = yx$ for all $x \in X$ and $y \in Y$, any $k \in K$ can be written as a word

$$k = a_1 \dots a_i b_1 \dots b_j$$

for some $a_1, \dots, a_i \in X^\pm$ and $b_1 \dots b_j \in Y^\pm$. Then

$$\hat{f}(k) = \hat{f}(a_1 \dots a_i b_1 \dots b_j) = (a_1, \dots, a_i, b_1 \dots b_j)$$

in $G \times H$. If $k \in \ker(\hat{f})$ then

$$\begin{aligned} a_1 \dots a_i &\in \langle R^{F(X)} \rangle \subseteq \langle R^{F(X \cup Y)} \rangle \subseteq \langle (R \cup S \cup [X, Y])^{F(X \cup Y)} \rangle \\ \text{and } b_1 \dots b_j &\in \langle S^{F(Y)} \rangle \subseteq \langle S^{F(X \cup Y)} \rangle \subseteq \langle (R \cup S \cup [X, Y])^{F(X \cup Y)} \rangle \end{aligned}$$

Since, by definition,

$$K = \langle X \cup Y : R \cup S \cup [X, Y] \rangle = F(X \cup Y) / \langle (R \cup S \cup [X, Y])^{F(X \cup Y)} \rangle,$$

both of these words $a_1 \dots a_i$ and $b_1 \dots b_j$ are equal to the identity e in K , and hence $k = e$ in K too. So the kernel $\ker(\hat{f})$ is trivial, hence $\hat{f}$ is injective and therefore an isomorphism. $\square$

In general, then, if we have a group G with presentation $\langle X : R \rangle$ and a group H with presentation $\langle Y : S \rangle$, we can write down a presentation for $G \times H$ by taking the union $X \cup Y$ of the generators and the union $R \cup S$ of the relators, and then forcing the generators of G to commute with the generators of H by means of the commutator $[X, Y]$.

¹⁰ Proposition 5.17, page 147.

If we skip this last step and don't require the generators of G to commute with the generators of H , we get a more general construct:

Definition 5.23 Let $G = \langle X : R \rangle$ and $H = \langle Y : S \rangle$. Then the **free product** of G and H , denoted $G * H$ is the group with presentation

$$G * H = \langle X \cup Y : R \cup S \rangle.$$

The free product $G * H$ is the most general possible group we can construct from G and H which still preserves G and H as subgroups. The next fact is fairly obvious, but worth noting.

Proposition 5.24 Let X and Y be sets, and let $F(X)$ and $F(Y)$ be the corresponding free groups. Then

$$F(X) * F(Y) = F(X \cup Y).$$

Similarly, if F_m and F_n are free groups of finite rank, then

$$F_m * F_n = F_{m+n}.$$

Proof This is an immediate consequence of Definition 5.23. We have $F(X) = \langle X : \rangle$ and $F(Y) = \langle Y : \rangle$, and therefore

$$F(X) * F(Y) = \langle X \cup Y : \rangle = F(X \cup Y).$$

The finite case $F_m * F_n = F_{m+n}$ follows by observing that $F_m = F(X)$ and $F_n = F(Y)$ for some sets X and Y with $|X| = m$ and $|Y| = n$. $\square$

Example 5.25 Let $G = \langle x : x^4 \rangle \cong \mathbb{Z}_4$ and $H = \langle y : y^6 \rangle \cong \mathbb{Z}_6$. Then the free product $G * H \cong \mathbb{Z}_4 * \mathbb{Z}_3$ is given by the presentation $\langle x, y : x^4 = y^6 \rangle$.

The groups $G \cong \mathbb{Z}_4$ and $H \cong \mathbb{Z}_6$ embed into $G * H$ by the homomorphisms $i: G \hookrightarrow G * H$ and $j: H \hookrightarrow G * H$ where $i(x) = x$ and $j(y) = y$.

Suppose now that both groups G and H have a subgroup which is isomorphic to some other group K . We can identify both copies of this subgroup and then take the free product, to get something called a **free product with amalgamation** or **amalgamated free product**.

Definition 5.26 Suppose that $G = \langle X : R \rangle$, $H = \langle Y : S \rangle$ and K are groups, and there exist injective homomorphisms $i: K \hookrightarrow G$ and $j: K \hookrightarrow H$. Let

$$T = \{i(k)j(k)^{-1} : k \in K\}.$$

Then the group

$$G *_K H = \langle X \cup Y : R \cup S \cup T \rangle$$

is the **free product** of G and H **amalgamated** over K .

This is the usual free product $G * H$ with additional relations which force $i(k) = j(k)$ for all $k \in K$.

Free products with amalgamation have useful applications in algebraic topology. In particular, an important result called the Seifert–van Kampen Theorem (or sometimes just van Kampen’s Theorem) tells us how to calculate a presentation for the fundamental group $\pi_1(X, *)$ of a topological space X in terms of the free product with amalgamation of the fundamental groups of its path-connected subspaces.

We won’t study free products or free products with amalgamation in detail, but here’s an illustrative example.

Example 5.27 Let $G = \mathbb{Z}_4 \cong \langle x : x^4 = e \rangle$ and $H = \mathbb{Z}_6 \cong \langle y : y^6 = e \rangle$. Each of these groups has a subgroup isomorphic to $\mathbb{Z}_2 = \langle t : t^2 = e \rangle$, which can be regarded as the image of the homomorphisms $i: \mathbb{Z}_2 \hookrightarrow \mathbb{Z}_4$ and $j: \mathbb{Z}_2 \hookrightarrow \mathbb{Z}_6$ given by

$$\begin{aligned} i(e) &= e, & i(t) &= x^2, \\ j(e) &= e, & j(t) &= y^3. \end{aligned}$$

From Example 5.25 we know that the free product $G * H$ is given by the presentation

$$G * H = \langle x, y : x^4 = y^6 = e \rangle.$$

To construct the free product amalgamated over $\mathbb{Z}_2$ via the homomorphisms i and j , we need to add in the extra relations we get from forcing $i(k) = j(k)$ for all $k \in \mathbb{Z}_2$. Setting $k = 1$ we get $i(1) = e = j(1)$, which doesn’t make any difference since it just gives us a trivial relation. Setting $k = t$, on the other hand, gives us the new relation $x^2 = y^3$, so

$$G *_{\mathbb{Z}_2} H = \mathbb{Z}_4 *_{\mathbb{Z}_2} \mathbb{Z}_6 = \langle x, y : x^4 = y^6 = e, x^2 = y^3 \rangle.$$

One of the first classes of nonabelian groups we met in Chapter 1 were the dihedral groups D_n . It turns out that these groups have presentations which follow a neat pattern.

Proposition 5.28 *Let D_n be the dihedral group of order $2n$ comprising the symmetries of a regular n -gon. Then*

$$D_n \cong \langle x, y : x^n = e, y^2 = e, yx = x^{-1}y \rangle.$$

Here, x corresponds geometrically to the $\frac{2\pi}{n}$ rotation, which obviously has order n , and y corresponds to one of the reflection, say m_1 . The third relation $yx = x^{-1}y$ says that doing a reflection followed by a rotation is the same as doing a rotation in the opposite direction and then reflecting. (Convince yourself that this is true.) What isn’t immediately clear is that these are the only relations we need for D_n .

Proof Any element of $G = \langle x, y : x^n = e, y^2 = e, yx = x^{-1}y \rangle$ can be written as a product of the generators x and y and their inverses x^{-1}

and y^{-1} . The relation $x^n = e$ means we can always write x^{-1} as x^{n-1} ; also x^{-k} can be rewritten as x^{n-k} for any $1 \leq k < n$. Similarly, $y^2 = e$ means we can replace any y^{-1} with y . Therefore we can always write any element of G as a product of positive powers of x and y , such that the maximum exponent of any occurrence of x is $n-1$ (since $x^n = e$) and the maximum exponent of any occurrence of y is 1 (since $y^2 = e$). The third relation $yx = x^{-1}y$ can be rewritten as $yx = x^{n-1}y$, so we can group all occurrences of y together (and simplify them to either e or a single y) and group all occurrences of x together. Thus any element of G is of the form x^k or yx^k for some k where $0 \leq k \leq n-1$.

This leaves us with precisely $2n$ words in G , namely

$$e, x, \dots, x^{n-1}, y, yx, \dots, yx^{n-1}$$

and hence $|G| = 2n = |D_n|$.

We still need to show that $G \cong D_n$ though, and to do this we have to consider the internal structure of dihedral groups in detail. Observe that the reflection m_k is equal to the product $m_1 r^{k-1}$ for $1 \leq k \leq n$.

So, define $f: G \rightarrow D_n$ such that $x^k \mapsto r^k$, and $yx^k \mapsto m_1 r^k = m_{k+1}$ for $0 \leq k \leq n-1$.

We claim that this is an isomorphism. It is obviously a bijection, so we need to check it's also a homomorphism, and to do that we have to verify the homomorphism condition for all four types of products:

$$\begin{aligned} f(x^k \cdot x^l) &= f(x^{k+l}) = r^{k+l} = r^k * r^l = f(r^k) * f(r^l) \\ f(x^k \cdot yx^l) &= f(yx^{n-k}x^l) = f(yx^{n-k+l}) = m_{n-k+l+1} \\ &= m_1 * r^{n-k+l} = m_1 * r^{n-k} * r^l = r^k * m_1 * r^l \\ &= r^k * m_{l+1} = f(x^k) * f(yx^l) \\ f(yx^k \cdot x^l) &= f(yx^{k+l}) = m_{k+l+1} = m_1 * r^{k+l} \\ &= m_1 * r^k * r^l = m_{k+1} * r^l = f(yx^k) * f(x^l) \\ f(yx^k \cdot yx^l) &= f(y^2 x^{n-k}x^l) = f(x^{n-k+l}) = r^{n-k+l} \\ &= m_1^2 * r^{n-k} * r^l = m_1 * r^k * m_1 * r^l \\ &= m_{k+1} * m_{l+1} = f(yx^k) * f(yx^l) \end{aligned}$$

Thus $f: G \rightarrow D_n$ is an isomorphism. $\square$

We can use this result to prove the following useful fact (promised in Chapter 2) about groups of order $2p$ where p is prime.

Proposition 5.29 *Let G be a group of order $2p$, where p is prime. Then G is isomorphic to either the cyclic group $\mathbb{Z}_{2p}$ or the dihedral group D_p .*

Proof By Proposition 2.34 we know that the order of any element $g \in G$ must divide $2p$. Hence $|g| = 1, 2, p$ or $2p$. By Proposition 1.19 we know that $|g| = 1$ if and only if g is the identity $e \in G$. Suppose,

then, that $g \neq e$ and hence $|g| = 2, p$ or $2p$.

If $|g| = 2p$ then $\langle g \rangle \cong \mathbb{Z}_{2p}$ comprises all of G , hence $G \cong \mathbb{Z}_{2p}$. We'll assume for the rest of the proof that G has no elements of order $2p$.

Suppose that there exists $g \in G$ such that $|g| = p$. Then the cyclic subgroup $\langle g \rangle = \{e, g, \dots, g^{p-1}\}$ has p elements and is therefore an index-2 subgroup of G . Choose some element $h \in G \setminus \langle g \rangle$. This element must have order 2 or p .

If $|h| = p$ then the number of elements in $\langle g \rangle \cup \langle h \rangle$ must be $2p-1$ since

$$|\langle g \rangle \cup \langle h \rangle| = |\langle g \rangle| + |\langle h \rangle| - |\langle g \rangle \cap \langle h \rangle| = p + p - 1 = 2p - 1.$$

Thus there must exist an element $k \in G$ which is in neither $\langle g \rangle$ or $\langle h \rangle$. The cyclic subgroup $\langle k \rangle$ must therefore consist of e and k only, since there are no other elements left. Hence $|k| = 2$.

If, on the other hand, $|h| = 2$ then $\langle h \rangle = \{e, h\}$, so set $k = h$. In either case we have found an element k of order 2 in $G \setminus \langle g \rangle$.

The subgroup $\langle g, k \rangle$ contains $\langle g \rangle$ as a proper subgroup and hence has index less than 2 in G . Since $\langle g \rangle$ has index 2 in G , then by Proposition 3.12 we know it is normal in G and therefore closed under conjugation. So $k * g * k^{-1} = g^m$ for some m .

Since $k^2 = e$ we can see that

$$\begin{aligned} g &= e * g * e = k^2 * g * k^{-2} = k * (k * g * k^{-1}) * k^{-1} = k * g^m * k^{-1} \\ &= (k * g * k^{-1})^m = (g^m)^m = g^{m^2} \end{aligned}$$

and so $g^{m^2-1} = e$, and therefore p must divide $m^2-1 = (m+1)(m-1)$. Hence p must divide either $m+1$ or $m-1$.

If p divides $m-1$ then $k * g * k^{-1} = g^m = g$, which means that $k * g = g * k$ and so G is abelian. If p is odd then $|k * g| = 2p$, which contradicts our assumption that G has no elements of order $2p$. The only other possibility is that $p = 2$, in which case $G \cong V_4 \cong D_2 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

If, however, p divides $m+1$ then $k * g * k^{-1} = g^m = g^{-1}$, which means that $k * g = g^{-1} * k$, and hence $G \cong D_p$ by Proposition 5.28. $\square$

We can now write presentations for all groups with up to seven elements:

$$\begin{array}{ll} 0 \cong \langle : \rangle & \mathbb{Z}_2 \cong \langle x : x^2 = e \rangle \\ \mathbb{Z}_3 \cong \langle x : x^3 = e \rangle & \mathbb{Z}_4 \cong \langle x : x^4 = e \rangle \\ V_4 \cong \langle x, y : x^2 = y^2 = e, xy = yx \rangle & \mathbb{Z}_5 \cong \langle x : x^5 = e \rangle \\ D_3 \cong \langle x, y : x^3 = y^2 = e, yx = x^{-1}y \rangle & \mathbb{Z}_6 \cong \langle x : x^6 = e \rangle \\ \mathbb{Z}_7 \cong \langle x : x^7 = e \rangle & \end{array}$$

There are five groups of order 8 (see Proposition 2.40), and we can

write down four of them straight away:

$$\mathbb{Z}_8 \cong \langle x : x^8 = e \rangle$$

$$\mathbb{Z}_2 \oplus \mathbb{Z}_4 \cong \langle x, y : x^2 = y^4 = e, xy = yx \rangle$$

$$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \cong \langle x, y, z : x^2 = y^2 = z^2 = e, xy = yx, xz = zx, yz = zy \rangle$$

$$D_4 \cong \langle x, y : x^4 = y^2 = e, yx = x^{-1}y \rangle$$

The remaining case, the quaternion group Q_8 , requires more work.

Proposition 5.30 *The quaternion group Q_8 has a presentation*

$$Q_8 \cong \langle x, y : x^4 = e, x^2 = y^2, xyx = y \rangle.$$

Proof Let $G = \langle x, y : x^4 = e, x^2 = y^2, xyx = y \rangle$. We want to show that this is indeed isomorphic to the quaternion group Q_8 , and we'll start by showing that it at least has the same number of elements.

The first relation $x^4 = e$ tells us that we need only consider positive powers of x , since we can replace any occurrence of x^{-1} with x^3 . The first and second relations together imply that $y^4 = e$, since

$$y^4 = (y^2)^2 = (x^2)^2 = x^4 = e.$$

Hence we only have to consider positive powers of y , since $y^{-1} = y^3$.

The third relation $xyx = y$ can be rearranged to $yx = x^{-1}y$, and then we can use the first relation to rewrite this as $yx = x^3y$.

Putting all this together, any finite word composed from the generators x and y and their inverses can be rearranged in the form $x^m y^n$ where m and n are non-negative integers. Using the second relation, we can rewrite y^n as either x^n (if n is even) or x^{n-1} (if n is odd) to get a word of the form x^k or $x^k y$. Since x has order 4 in G we know that $0 \leq k < 4$, and so G has eight elements:

$$G = \{e, x, x^2, x^3, y, xy, x^2y, x^3y\}.$$

The multiplication table for this group is shown in Table 5.2 along with that for Q_8 , reordered to make the structural similarities clearer.

The function $f: G \rightarrow Q_8$ that maps

$$\begin{array}{llll} e \mapsto E & x \mapsto I & x^2 \mapsto -E & x^3 \mapsto -I \\ y \mapsto J & xy \mapsto K & x^2y \mapsto -J & x^3y \mapsto -K \end{array}$$

is the required isomorphism. $\square$

The quaternion group is the first in a sequence of nonabelian groups of order $4n$ for some positive integer n .

Example 5.31 The **dicyclic group** Dic_n of order $4n$ is defined by the presentation

$$Dic_n \cong \langle x, y : x^{2n} = e, x^n = y^2, xyx = y \rangle.$$

G	e	x	x^2	x^3	y	xy	x^2y	x^3y
e	e	x	x^2	x^3	y	xy	x^2y	x^3y
x	x	x^2	x^3	e	xy	x^2y	x^3y	y
x^2	x^2	x^3	e	x	x^2y	x^3y	y	xy
x^3	x^3	e	x	x^2	x^3y	y	xy	x^2y
y	y	x^3y	x^2y	xy	x^2	x	e	x^3
xy	xy	y	x^3y	x^2y	x^3	x^2	x	e
x^2y	x^2y	xy	y	x^3y	e	x^3	x^2	x
x^3y	x^3y	x^2y	xy	y	x	e	x^3	x^2

Q_8	E	I	$-E$	$-I$	J	K	$-J$	$-K$
E	E	I	$-E$	$-I$	J	K	$-J$	$-K$
I	I	$-E$	$-I$	E	K	$-J$	$-K$	J
$-E$	$-E$	$-I$	E	I	$-J$	$-K$	J	K
$-I$	$-I$	E	I	$-E$	$-K$	J	K	$-J$
J	J	$-K$	$-J$	K	$-E$	I	E	$-I$
K	K	J	$-K$	$-J$	$-I$	$-E$	I	E
$-J$	$-J$	K	J	$-K$	E	$-I$	$-E$	I
$-K$	$-K$	$-J$	K	J	I	E	$-I$	$-E$

Table 5.2: The multiplication tables for the group

$$G \cong \langle x, y : x^4 = e, x^2 = y^2, xyx = y \rangle$$

and the quaternion group Q_8

The case $n = 1$ yields

$$\text{Dic}_1 \cong \langle x, y : x^2 = e, x = y^2, xyx = y \rangle \cong \mathbb{Z}_4$$

and the group Dic_2 is isomorphic to the quaternion group Q_8 as discussed in Proposition 5.30.

All of the examples we've seen so far have been of finitely-presented groups. Here is an important example of a group which doesn't have a finite presentation.

Proposition 5.32 *Let $\mathbb{Q}$ denote the additive group of rational numbers. Then $\mathbb{Q}$ is not finitely presented; in fact*

$$\mathbb{Q} \cong \langle x_1, x_2, \dots : x_n^n = x_{n-1} \text{ for } n > 1 \rangle.$$

Proof To see that $\mathbb{Q}$ is not finitely presented, consider some finite subset $X \subseteq \mathbb{Q}$ and let n be the lowest common multiple of the denominators of all the elements in X . Now consider the subgroup $\langle X \rangle \leq \mathbb{Q}$ generated additively by all of the elements of X . Since all of the denominators of elements in X divide n , this is also true for the denominators of elements in the subgroup $\langle X \rangle$, and so X can't generate all of $\mathbb{Q}$, since we'll always be able to find some element of $\mathbb{Q}$ whose denominator doesn't divide n .

It so happens that $\mathbb{Q}$ can be generated by the infinite set $\{\frac{1}{n!} : n \in \mathbb{N}\}$, since any rational number $\frac{m}{n}$ can be written as

$$\frac{m}{n} = \frac{m(n-1)!}{n!} = m(n-1)! \frac{1}{n!}.$$

Let $X = \{x_1, x_2, \dots\}$ and let $G = \langle x_1, x_2, \dots : nx_n = x_{n-1} \text{ for } n > 1 \rangle$. Define a function $f: X \rightarrow \mathbb{Q}$ by $f(x_n) = \frac{1}{n!}$. Then $f(x_n^n) = \frac{n}{n!} = \frac{1}{(n-1)!} = f(x_{n-1})$, so by Proposition 5.17 we know that f extends to a homomorphism $\hat{f}: G \rightarrow \mathbb{Q}$. This homomorphism is surjective since $\mathbb{Q}$ is generated by the set $\{\frac{1}{n!} : n \in \mathbb{N}\}$.

The group G is abelian, since the relations $x_n = x_{n-1}^n$ tell us each generator x_n can be written as a power of any of the previous $n-1$ generators $x_1, \dots, x_{n-1}$, and hence any two generators x_i and x_j commute, which means any two elements of G commute, so G is abelian. So, any element of G can be written as a finite product $x_1^{k_1} x_2^{k_2} \dots x_n^{k_n}$ for some integers $k_1, \dots, k_n \in \mathbb{Z}$. Since $x_n^n = x_{n-1}$ we can use the Euclidean algorithm to write $k_n = qn + r$ where $0 \leq r < n$. So

$$x_n^{k_n} = x_n^{qn+r} = x_n^{qn} x_n^r = x_{n-1}^q x_n^r$$

which tells us that $0 \leq k_n < n$. Repeating this process for the generators $x_{n-1}, x_{n-2}, \dots, x_2$ lets us rewrite the word in a canonical form

$$x_n^{m_1} x_2^{m_2} \dots x_n^{m_n}$$

where $0 \leq m_i < i$ for $2 \leq i \leq n$.

We need to show that $\hat{f}$ is injective, which we can do by showing its kernel is trivial. Let $g \in \ker \hat{f}$ be some element of this kernel. By the above discussion we know $g = x_1^{m_1} x_2^{m_2} \dots x_n^{m_n}$ for some integers $m_1, \dots, m_n \in \mathbb{Z}$ such that $0 \leq m_i < i$ for all $i > 1$. Looking at the image $\hat{f}(g)$ of this element, we see

$$\hat{f}(g) = \frac{m_1}{1!} + \frac{m_2}{2!} + \dots + \frac{m_n}{n!} = 0$$

Multiplying by $(n-1)!$ we get

$$(n-1)! \left(\frac{m_1}{1!} + \frac{m_2}{2!} + \dots + \frac{m_n}{n!} \right) = m + \frac{m_n}{n} = 0$$

for some integer $m \in \mathbb{Z}$. This means that $\frac{m_n}{n} = -m$ must also be an integer, but we know already that $0 \leq m_n < n$ and so the only possibility is that $m_n = 0$. Therefore

$$\hat{f}(g) = \frac{m_1}{1!} + \frac{m_2}{2!} + \dots + \frac{m_{n-1}}{(n-1)!} + 0 = 0$$

and we can repeat the same process with $(n-2)!$ to show that $m_{n-1} = 0$, and then repeat $(n-3)!$ more times to show that $m_2 = \dots = m_n = 0$.

Hence $g = x_1^{m_1}$ and since $\hat{f}(g) = \hat{f}(x_1^{m_1}) = m_1 = 0$ this tells us that $g = x_1^0 = e$ and so $\ker(\hat{f})$ is trivial. Therefore $\hat{f}$ is injective, and since we've already shown it's surjective that means it's an isomorphism, and so $Q \cong \langle x_1, x_2, \dots : x_n^n = x_{n-1} \text{ for } n > 1 \rangle$ as claimed. $\square$

Another important class of finite groups we met in Chapter 1 are the symmetric groups S_n , and with a little bit of work we can find presentations for these as well.

Proposition 5.33 *The symmetric group S_n may be defined by the presentation*

$$S_n \cong \left\langle x_1, \dots, x_{n-1} : \begin{array}{ll} x_i^2 = e & 1 \leq i \leq n-1 \\ (x_i x_{i+1})^3 = e & 1 \leq i \leq n-2 \\ x_i x_j = x_j x_i & 1 \leq i < j \leq n-1, \quad j-i > 1 \end{array} \right\rangle.$$

Proof Let G_n be the group defined by the above presentation. We aim to construct an isomorphism from G_n to S_n .

This presentation is inspired by Proposition 1.60(ii), which says that S_n is generated by transpositions of adjacent objects; that is, transpositions of the form $x_i := (ii+1)$ for $1 \leq i \leq n-1$. We can easily verify that all three types of relations are satisfied by these transpositions. The first follows immediately from the fact that transpositions have order 2:

$$x_i^2 = (ii+1)^2 = (ii+1)(ii+1) = \iota.$$

The second follows by a straightforward calculation:

$$\begin{aligned} (x_i x_{i+1})^3 &= ((ii+1)(i+1i+2))^3 = \\ &= (ii+1)(i+1i+2)(ii+1)(i+1i+2)(ii+1)(i+1i+2) = \iota \end{aligned}$$

The third follows from the fact that if $|i - j| > 1$ then the transpositions $x_i = (ii+1)$ and $x_j = (jj+1)$ are disjoint, and therefore commute:

$$(ii+1)(jj+1) = (jj+1)(ii+1).$$

We can thus define a surjective homomorphism $f: G_n \rightarrow S_n$ by mapping each generator x_i in G_n to the transposition $(ii+1)$ in S_n . If we can also show this homomorphism is injective, then we've finished.

To do this, we can use the fact that $|S_n| = n!$ and show by induction that $|G_n| \leq n!$, which will then imply that f must be injective.

When $n = 1$ we have the trivial group $G_1 = \langle : \rangle$, which has order $|G_1| = 1 \leq 1!$ as required. For $n = 2$ we have $G_2 = \langle x_1 : x_1^2 = e \rangle \cong \mathbb{Z}_2$ and hence $|G_2| = 2 \leq 2!$ as well.

Now suppose $n \geq 3$ and $|G_{n-1}| \leq (n-1)!$ for the induction hypothesis. Let $H \leq G_n$ be the subgroup generated by the first $n-2$ generators $x_1, \dots, x_{n-2}$ and define

$$y_0 = e \quad \text{and} \quad y_i = x_{n-1}x_{n-2} \dots x_{n-i} \quad \text{for } 1 \leq i \leq n-1.$$

So

$$y_0 = e, \quad y_1 = x_{n-1}, \quad y_2 = x_{n-1}x_{n-2}, \quad \dots, \quad y_{n-1} = x_{n-1}x_{n-2} \dots x_1.$$

Now let

$$A = \{hy_i : h \in H \text{ and } 0 \leq i \leq n-1\}.$$

This is certainly a subset of G_n , but we want to show that it's equal to the whole of G_n . To do this, we consider products of the form hy_ix_j , where h is some element of H . There are six cases to consider:

Case 1 ($i = 0, j < n-1$) $hy_ix_j = hy_0x_j = hx_j \in Hy_0 \subseteq A$.

Case 2 ($i = 0, j = n-1$) $hy_ix_j = hy_0x_{n-1} = hx_{n-1} = hy_1 \in Hy_1 \subseteq A$.

Case 3 ($i > 0, j > n-i$)

$$\begin{aligned} hy_ix_j &= h(x_{n-1} \dots x_j x_{j-1} \dots x_{n-i})x_j \\ &= h(x_{n-1} \dots x_j x_{j-1} x_j x_{j-2} \dots x_{n-i}) \\ &= h(x_{n-1} \dots x_{j+1} x_{j-1} x_j x_{j-1} x_{j-2} \dots x_{n-i}) \\ &= h(x_{j-1} x_{n-1} \dots x_{j+1} x_j x_{j-1} \dots x_{n-i}) \\ &= hx_{j-1}y_i \in Hy_i \subseteq A. \end{aligned}$$

Case 4 ($i > 0, j = n-i$)

$$\begin{aligned} hy_ix_j &= hx_{n-1} \dots x_{n-i}x_{n-i} \\ &= hx_{n-1} \dots x_{n-i+1} \\ &= hy_{i-1} \in Hy_{i-1} \subseteq A. \end{aligned}$$

Case 5 ($i > 0, j = n-i-1$)

$$\begin{aligned} hy_ix_j &= hx_{n-1} \dots x_{n-i}x_{n-i-1} \\ &= hy_{i+1} \in Hy_{i+1} \subseteq A. \end{aligned}$$

Case 6 ($i > 0, j < n-i-1$) $hy_ix_j = hx_jy_i \in Hy_i \subseteq A$.

Therefore $Ax_j \subseteq A$ for all j , and since $x_j^2 = e$ we also know that $Ax_j^{-1} = Ax_j \subseteq A$.

Not only that, if w is any word in the generators $x_1, \dots, x_{n-1}$ and their inverses, then $Aw \subseteq A$. (This can be rigorously proved by induction on the length $l(w)$ of the word w , but we'll skip the details here.)

In fact, any word $w \in G_n$ is equal to some element of A , because $e \in A$ and so $w = ew \in Aw \subseteq A$. Hence $A = G_n$.

The relations in G_n involving just the first $n-2$ generators $x_1, \dots, x_{n-2}$ are exactly those which lie in G_{n-1} , so the map $\phi: G_{n-1} \rightarrow G_n$ such that each $x_i \mapsto x_i$ is a homomorphism with $\text{im } \phi = H$. Then

$$|G_n| \leq |A| \leq n|H| \leq n(n-1)! = n!$$

Since $|G_n| \leq |A|$ and $f: G_n \rightarrow S_n$ is surjective, and since we know from Proposition 1.52 that $|S_n| = n!$ this means that the only possibility is that f is also injective, and hence an isomorphism.

Thus $G_n \cong S_n$ and the given presentation is a valid one for S_n . $\square$

The given presentation for S_n is a bit complicated in general, so it's illuminating to look at a specific example.

Example 5.34 The symmetric group S_3 has presentation

$$\langle x, y : x^2 = y^2 = (xy)^3 = e \rangle.$$

The symmetric group S_4 has presentation

$$\langle x, y, z : x^2 = y^2 = z^2 = (xy)^3 = (yz)^3 = e, \quad xz = zx \rangle.$$

We also happen to know $S_3 \cong D_3$, and Proposition 5.28 tells us that

$$D_3 \cong \langle a, b : a^3 = b^2 = e, ba = a^{-1}b \rangle.$$

Since these two presentations yield isomorphic groups, there must surely be some structural connection between them. Ideally, we want a precise way of deciding whether two presentations define isomorphic groups or not. As with many questions in mathematics, the answer is simple and elegant, but sometimes not as practically useful as we might prefer. First, the good news.

Proposition 5.35 Let $G \cong \langle X : R \rangle$ be a group.

- (i) If $r \in \langle R^{F(X)} \rangle$ is a relator of G then $G \cong \langle X : R \cup \{r\} \rangle$.
- (ii) Let $y \notin X$ be another generator and $w \in W(X^\pm)$ be some word in $F(X)$. Then $G \cong \langle X \cup \{y\} : R \cup \{yw^{-1}\} \rangle$, where the isomorphism restricts to the identity id_X on X .

Proof To show part (i) we note that if r is a relator of G , then it's in the normal closure $\langle R^{F(X)} \rangle$, and hence $\langle (R \cup \{r\})^{F(X)} \rangle = \langle R^{F(X)} \rangle$.

Therefore

$$\langle X : R \cup \{r\} \rangle = F(X) / \langle (R \cup \{r\})^{F(X)} \rangle = F(X) / \langle R^{F(X)} \rangle = \langle X : R \rangle \cong G.$$

¹¹ Proposition 5.17, page 147.

The proof of part (ii) is a little more involved, and uses the fundamental mapping property of group presentations.¹¹ Let $Y = X \cup y$ and as usual let $F(Y)$ denote the free group on Y . Let $H = \langle X \cup \{y\} : R \cup \{yw^{-1}\} \rangle$ be the group obtained by adding the new generator y and the new relator yw^{-1} .

We want to show that $G \cong H$, so to that end we define a function $\phi: X \rightarrow H$ such that every generator in X gets mapped to itself in $Y = X \cup y \subset H$. Every relator in G is also a relator in H , so we can use Proposition 5.17 to extend this to a homomorphism $\hat{\phi}: G \rightarrow H$.

We can also define a homomorphism going the other way: let $\psi: Y \rightarrow G$ be such that $\psi(x) = x$ for all $x \in X$ and $\psi(y) = w$. For any relator $r \in R$ we have $\psi(r) = e_G$, and also $\psi(yw^{-1}) = \psi(y)w^{-1} = ww^{-1} = e_G$. Hence we can use Proposition 5.17 again to extend this to a homomorphism $\hat{\psi}: H \rightarrow G$.

The composite $\hat{\psi} \circ \hat{\phi}: G \rightarrow G$ extends the inclusion map $i: X \hookrightarrow G$, and $\hat{\phi} \circ \hat{\psi}: H \rightarrow H$ extends the inclusion map $j: Y \hookrightarrow H$. The identity homomorphisms $\text{id}_G: G \rightarrow G$ and $\text{id}_H: H \rightarrow H$ also do this, and so by the uniqueness property in Proposition 5.17, $\hat{\psi} \circ \hat{\phi} = \text{id}_G$ and $\hat{\phi} \circ \hat{\psi} = \text{id}_H$, so $\hat{\phi}$ and $\hat{\psi}$ must be isomorphisms. Hence $G \cong H$. $\square$

This tells us that there are two specific things we can do to a group presentation that won't fundamentally change the group it describes. We can add a redundant relator, and we can add a new generator together with a new relator that equates it to an existing element of the group. We can also do these in reverse: we can delete a redundant relator, and we can remove a redundant generator and corresponding relator. More precisely, we have the following definition:

Definition 5.36 Let $G = \langle X : R \rangle$. By Proposition 5.35 the following operations will give a presentation for a group isomorphic to G :

- X_+ Add a new generator $y \notin X$ together with a new relator yw^{-1} where $w \in W(X^\pm)$. Then $G \cong \langle X \cup \{y\} : R \cup \{yw^{-1}\} \rangle$.
- X_- Delete a redundant generator $y \in X$ and relator $yw^{-1} \in R$ if neither w nor any other relator in R contains y or y^{-1} . Then $G \cong \langle X \setminus \{y\} : R \setminus \{yw^{-1}\} \rangle$.
- R_+ Add a new relator $r \in \langle R^{F(X)} \rangle \setminus R$. Then $G \cong \langle X : R \cup \{r\} \rangle$.
- R_- Delete a redundant relator $r \in \langle R^{F(X)} \rangle$ where $\langle (R \setminus \{r\})^{F(X)} \rangle = \langle R^{F(X)} \rangle$. Then $G \cong \langle X : R \setminus \{r\} \rangle$.

These operations are called **Tietze transformations**, originally formulated by the Austrian mathematician Heinrich Tietze (1880–1964).

Tietze transformations are typically applied in combinations, and are best illustrated by a few concrete examples.

Example 5.37 Let $G_{m,n} = \langle X : R \rangle = \langle x, y : x^m y^{-n} \rangle$. We can apply an R_+ operation to introduce the relator $y^n x^{-m}$. This is certainly allowed, since if $x^m y^{-n}$ is in the normal closure $\langle R^{F(X)} \rangle$ then so is $y^n x^{-m} = (x^m y^{-n})^{-1}$, so by Proposition 5.35 we have

$$G_{m,n} = \langle X : R \cup \{y^n x^{-m}\} \rangle = \langle x, y : x^m y^{-n}, y^n x^{-m} \rangle.$$

Having done this, we can use R_- to delete $x^m y^{-n}$, since the normal closure $\langle (R \setminus \{x^m y^{-n}\})^{F(X)} \rangle = \langle R^{F(X)} \rangle$. This gives us

$$\langle x, y : y^n x^{-m} \rangle$$

and again by Proposition 5.35 this will be the same group $G_{m,n}$. We will quite often have cause to use a double operation of this type, where a relator r is replaced with an equivalent relator s . We may denote such an addition–deletion move as $R_\pm$.

(This group $G_{m,n}$ happens to be the fundamental group of the complement of a torus knot of type (m, n) .)

We can take this a little further and replace any relator with a conjugate of itself, in particular a cyclic permutation:

Example 5.38 Let $G_{m,n} = \langle X : R \rangle = \langle x, y : x^m y^{-n} \rangle$. Since $x^m y^{-n} = e$ we can take the conjugate with x to get $x^{-1} x^m y^{-n} x = x^{-1} e x = e$ and hence $x^{m-1} y^{-n} x = e$. So $x^{m-1} y^{-n} x$ is also in the normal closure $\langle R^{F(X)} \rangle$ and hence we can perform an $R_\pm$ operation to get

$$G_{m,n} = \langle x, y : x^{m-1} y^{-n} x \rangle.$$

More generally, we can replace any relator by a conjugate or cyclic permutation by means of the $R_\pm$ double operation.

Another consequence of this example relates to the X_- move, in which a generator y can be deleted if there exists a relator of the form yw^{-1} where w is some word in the remaining generators and their inverses, and no other relator contains y . Example 5.38 tells us it doesn't matter if the chosen generator y isn't at the beginning of some relator yw^{-1} , because we can use a sequence of $R_\pm$ moves to cyclically permute a relator $u^{-1} y v^{-1}$ into the form $y v^{-1} u^{-1} = y(uv)^{-1}$.

The next example shows that it doesn't even matter if y , the generator we want to eliminate, happens to occur in other relators as well, as long as there's one that just contains a single occurrence of y or y^{-1} .

Example 5.39 Let $G = \langle X : R \rangle = \langle x, y : x^2 y^{-3}, y x^5 \rangle$. We can take the relator $y x^5 = e$ and rearrange it to get $y = x^{-5}$, which we can then substitute into the other relator to get $e = x^2 (x^{-5})^{-3} = x^{17}$.

This means that $x^{17} \in \langle R^{F(X)} \rangle$ and so we can perform an R_+ move to obtain the presentation $G = \langle X : R \cup \{x^{17}\} \rangle = \langle x, y : x^2y^{-3}, yx^5, x^{17} \rangle$.

We can now apply an R_- move to delete the now redundant relator x^2y^{-3} , obtaining the presentation

$$G = \langle x, y : yx^5, x^{17} \rangle.$$

Finally, we can perform an X_- transformation to delete the generator y and the relator yx^5 , obtaining the much simpler presentation

$$G = \langle x : x^{17} \rangle \cong \mathbb{Z}_{17}.$$

We will denote a compound transformation of this type, consisting of a number of R_+ moves followed by a corresponding number of R_- moves and an X_- move, by the symbol RX_- .

The next example leads on from Example 5.34 and Proposition 5.28 to show that the two presentations for $D_3 \cong S_3$ are indeed equivalent.

Example 5.40 From Proposition 5.28 we know that

$$D_3 \cong \langle a, b : a^3, b^2, (ab)^2 \rangle,$$

and from Proposition 5.33 and Example 5.34 we know that

$$D_3 \cong S_3 \cong \langle x, y : x^2, y^2, (xy)^3 \rangle.$$

First we apply an X_+ move to the first presentation, introducing the generator x and the relator $x^{-1}ab$ (equivalent to $x = ab$). This yields

$$\langle a, b, x : a^3, b^2, (ab)^2, x^{-1}ab \rangle.$$

Next, we perform another X_+ move to introduce another generator y and relator by^{-1} (which is equivalent to the relation $y = b$). This gives

$$\langle a, b, x, y : a^3, b^2, (ab)^2, x^{-1}ab, by^{-1} \rangle.$$

We can now use an R_+ move to add a new relator y^2 and follow it with an R_- move to delete the relator b^2 ; this results in the presentation

$$\langle a, b, x, y : a^3, (ab)^2, x^{-1}ab, by^{-1}, y^2 \rangle.$$

Similarly, we can replace the relator $x^{-1}ab$ with $x^{-1}ay$, and replace the relator $(ab)^2$ with $(ay)^2$ to obtain the presentation

$$\langle a, b, x, y : a^3, (ay)^2, x^{-1}ay, by^{-1}, y^2 \rangle.$$

We can now perform an X_- move to delete the generator b and the relator by^{-1} , in order to get the presentation

$$\langle a, x, y : a^3, (ay)^2, x^{-1}ay, y^2 \rangle.$$

Another $R_{\pm}$ transformation replaces the relator $x^{-1}ay$ with its cyclic conjugate ayx^{-1} , and then we perform two more $R_{\pm}$ moves to replace

a^3 with $(xy^{-1})^3$ and replace $(ay)^2$ with x^2 to get the presentation

$$\langle a, x, y : (xy^{-1})^3, x^2, ayx^{-1}, y^2 \rangle.$$

All that remains now is to do another X_- transformation to delete the generator a and the relator ayx^{-1} , and then do another $R_{\pm}$ transformation to replace $(xy^{-1})^3$ with $(xy)^3$ to get the presentation into the required form

$$\langle x, y : (xy)^3, x^2, y^2 \rangle.$$

This example is a bit long-winded: we went through the sequence of Tietze transformations in a bit more detail than would ordinarily be required, but hopefully it's given you a clearer idea of how the process works.

In a sense this process is analogous to the study of systems of simultaneous linear equations. Given a collection of such equations, we want to find the solution set; that is, the complete set of possible values each of the variables can take and still satisfy the equation. There are certain basic operations we can perform on a given system that won't change the solution set, but which will (if we've done it all correctly) simplify the system and make it easier to solve. We can multiply both sides of an equation by a nonzero scalar, and we can add a scalar multiple of one equation to another. We'll come back to this idea in the next section when we study presentations of abelian groups, where the similarities will become more obvious.

Proposition 5.35 tells us that the Tietze transformations X_+ , X_- , R_+ and R_- don't change the group in question. That is, if we have a group G defined by a presentation $\langle X : R \rangle$ and we perform a finite sequence of Tietze transformations on it, ending up with a new presentation $\langle Y : S \rangle$, then this new presentation defines a group which is isomorphic to the original group G .

What Proposition 5.35 doesn't tell us is whether, given two presentations $\langle X : R \rangle$ and $\langle Y : S \rangle$ for the same group, we can transform one into another by a finite sequence of Tietze transformations. As it happens, we can do exactly this, as the following proposition confirms.

Proposition 5.41 *Let $\langle X : R \rangle$ and $\langle Y : S \rangle$ be two presentations for the same group G . Then there exists a finite sequence of Tietze moves transforming $\langle X : R \rangle$ into $\langle Y : S \rangle$.*

Proof Given that $G \cong \langle X : R \rangle \cong \langle Y : S \rangle$ it follows that both X and Y generate G , and hence we can write every element of X as a word in $F(Y)$, and write every element of Y as a word in $F(X)$. Denote these sets of equations by $X(Y)$ and $Y(X)$ respectively. Starting with the presentation $\langle X : R \rangle$, we perform X_+ and X_- transformations to

obtain the presentation

$$\langle X, Y : R(X) \cup Y(X) \rangle$$

(where $R(X)$ denotes the relators R expressed as words in $F(X)$). Next perform R_+ moves to transform this into the presentation

$$\langle X, Y : R(X) \cup Y(X) \cup X(Y) \rangle.$$

We can now use RX_- moves to eliminate the generators in X to get

$$\langle Y : R(X(Y)) \cup Y(X(Y)) \rangle.$$

Here $R(X(Y))$ denotes the relators from R written in terms of the generators in Y via the equations in $X(Y)$, and $Y(X(Y))$ denotes the generators in Y written as words in $F(Y)$ via the equations in $X(Y)$. A sequence of R_+ moves allows us to adjoin the relators $S(Y)$ to get

$$\langle Y : R(X(Y)) \cup Y(X(Y)) \cup S(Y) \rangle$$

and finally we perform a sequence of R_- moves to delete the relators in $R(X(Y))$ and $Y(X(Y))$ to obtain the presentation

$$\langle Y : S \rangle$$

as required. □

Putting Propositions 5.35 and 5.41 together, we now know that two presentations $\langle X : R \rangle$ and $\langle Y : S \rangle$ give isomorphic groups if and only if we can turn one into another by a finite sequence of the four Tietze transformations (which in practice we usually employ in compound forms such as $X_\pm$ and RX_-).

This is the good news. The bad news is that, in general, whether or not two arbitrary words yield the same element of a given group is a formally undecidable question: this is known as the **word problem**. There are many classes of groups for which the word problem is solvable, including finite groups, the Euclidean groups we met in Definitions 1.47 and 1.48, and the Coxeter groups and braid groups we'll meet in Section 5.C.

But in general the word problem is unsolvable, and one of the consequences of this is that there is no systematic, algorithmic method for deciding whether two arbitrary presentations determine isomorphic groups. (The proof of Proposition 5.41 worked because we'd decided from the start that the two presentations were isomorphic.)

Nevertheless, this machinery of group presentations is still very useful for many purposes, and there are many powerful techniques and results that can be derived using it. In the next section we will state and prove a classification theorem for finitely generated abelian groups, and in Section 5.A we will study the Todd–Coxeter algorithm which calculates several useful pieces of information about a given group.

5.3 Finitely generated abelian groups

IN SECTION 5.1 WE STUDIED FREE GROUPS, the most general possible groups we can generate from a particular set of generators. It's worth asking what is the most general possible *abelian* group we can construct from a given set of generators. We could go through the same construction as for free groups, in addition requiring that all the generators commute with each other. This definitely results in an abelian group whose elements are commuting words in the generators. Unlike with the more general case, we allow arbitrary arbitrary reordering of generators, which in particular means that we can group together generators of the same type in a nice, neat and straightforward way.

For example, suppose $X = \{x, y, z\}$ is a set of generators, and that

$$w_1 = x^2y^{-1}xz^3y^{-2} \quad \text{and} \quad w_2 = y^{-1}z^{-1}x^2z^3y^{-2}.$$

Allowing arbitrary reordering means we can reduce these two words to get

$$w_1 = x^3y^{-3}z^3 \quad \text{and} \quad w_2 = x^2y^{-3}z^2.$$

Concatenation, reduction and reordering of these words gives

$$w_1w_2 = x^5y^{-6}z^5 = w_2w_1.$$

As before, this concatenation–reordering–reduction operation is an associative binary operation on the set of commuting words in $X^\pm$, the empty word serves as an identity element, and each word has a unique, well-defined inverse; for example

$$w_1^{-1} = x^{-3}y^3z^{-3} \quad \text{and} \quad w_2^{-1} = x^{-2}y^3z^{-2}.$$

We therefore have all the necessary ingredients for an abelian group:

Definition 5.42 The group just described is the **free abelian group** $FA(X)$ generated by the set X . If X is finite and $|X| = n$, we will often denote this as FA_n .

Free abelian groups $FA(X)$ satisfy a similar fundamental property as the free group $F(X)$ except we restrict our attention to abelian groups:

Definition 5.43 Let $X \subset F$ be a set, and A be some arbitrary abelian group. Then an abelian group F is **free abelian** on X if any function $f: X \rightarrow A$ extends uniquely to a homomorphism $\tilde{f}: F \rightarrow A$.

A slight modification of the proof of Proposition 5.6 confirms that these two definitions are equivalent. But another way to look at this, and one that will lead us naturally into the next section, is to say that in constructing the free abelian group $FA(X)$ essentially what we've done is to take $F(X)$ and make it abelian: we've made everything commute with everything else.

It is the peculiar beauty of this method, gentlemen, and one which endears it to the really scientific mind, that under no circumstance can it be of the smallest possible utility.

— Henry Smith (1826–1883)



Wikimedia Commons

(Julius Wilhelm) Richard Dedekind (1831–1916) was born in Braunschweig, Germany, the fourth and youngest child of Julius Dedekind, a professor at the Collegium Carolinum (now the Technische Universität Braunschweig). He entered the University of Göttingen in 1850, where he studied with Carl Friedrich Gauss (1777–1855). Dedekind became Gauss' last student, and in 1852 was awarded a doctorate for a thesis entitled *Über die Theorie der Eulerschen Integrale* ("On the theory of Eulerian integrals"). As is often the case, this thesis was adequate but did not display the talent evident in his later work. Over the next two years he studied the latest developments in mathematics, and in 1854 was awarded his *habilitation* (a postdoctoral qualification required to teach at a German university).

During his career he made many important advances in analysis, algebra and number theory. In particular, he devised the notion of a **Dedekind cut**, used in one construction of the real numbers $\mathbb{R}$. He became an early admirer of the work of Georg Cantor (1845–1918) on transfinite numbers, and was a staunch ally in his disputes with Leopold Kronecker (1823–1891).

He also devised the concept of an **ideal** and used it to provide a purely algebraic proof of the Riemann–Roch Theorem, an important result in algebraic geometry.

He retired in 1894 and returned to Braunschweig, but continued his research and taught occasionally for some years. He died in 1916 aged 84.

This is exactly the **abelianisation** process we met in Chapter 3. In constructing $FA(X)$ we've factored $F(X)$ by its commutator subgroup $[F(X), F(X)]$. That is, $FA(X) = F(X)^{\text{ab}}$.

Depending on whether we've decided to use multiplicative or additive notation, we can write a typical element w of $FA(X)$ as either

$$w = x_1^{n_1} x_2^{n_2} \dots = \prod_{x \in X} x^{n_x} \quad \text{or} \quad w = n_1 x_1 + n_2 x_2 + \dots = \sum_{x \in X} n_x x$$

where $n_1, \dots \in \mathbb{Z}$. Most of the time, we'll use additive notation for free abelian groups and their quotients.

Proposition 5.44 *Let $X = \{x_1, \dots, x_n\}$ be a finite set of generators. Then the free abelian group $FA_n = FA(X)$ is isomorphic to the n -fold direct sum $\mathbb{Z}^n = \mathbb{Z} \oplus \dots \oplus \mathbb{Z}$.*

Proof Written additively, an element of FA_n is a word of the form

$$w = m_1 x_1 + \dots + m_n x_n$$

where $m_1, \dots, m_n \in \mathbb{Z}$.

The function $f: FA_n \rightarrow \mathbb{Z}^n$ that maps a word w to the n -tuple $(m_1, \dots, m_n)$ in $\mathbb{Z}^n$ is clearly bijective. It's also a homomorphism, since for any

$$u = a_1 x_1 + \dots + a_n x_n \quad \text{and} \quad v = b_1 x_1 + \dots + b_n x_n$$

we have $uv = (a_1 + b_1)x_1 + \dots + (a_n + b_n)x_n$ and so

$$f(uv) = (a_1 + b_1, \dots, a_n + b_n) = (a_1, \dots, a_n) + (b_1, \dots, b_n) = f(u) + f(v)$$

as required. Hence f is an isomorphism. $\square$

The following proposition is the free abelian analogue of a very similar fact in linear algebra.

Proposition 5.45 *Let F be an abelian group, and let $X = \{x_1, \dots, x_n\} \subseteq F$. Then F is free abelian of rank n and the elements of X are $\mathbb{Z}$ -linearly independent if and only if any element of F can be written as a unique $\mathbb{Z}$ -linear combination of the elements $x_1, \dots, x_n$.*

Proof Suppose that any element $g \in F$ can be written uniquely as the $\mathbb{Z}$ -linear combination

$$g = a_1 x_1 + \dots + a_n x_n$$

where $a_1, \dots, a_n \in \mathbb{Z}$. Clearly X generates F , but we want to show that F is actually the free abelian group $FA(X)$ generated by X , and which consists of all possible formal $\mathbb{Z}$ -linear combinations of $x_1, \dots, x_n$.

Let $f: X \hookrightarrow F$ be the inclusion function. Then by the fundamental mapping property of free abelian groups, there exists a unique homomorphism $\tilde{f}: FA(X) \rightarrow F$ which extends f , and which maps a given $\mathbb{Z}$ -linear combination in $FA(X)$ to the corresponding sum in F . We

want to show that this is an isomorphism. It is clearly surjective, since we've already hypothesised that any element of F can be expressed as a $\mathbb{Z}$ -linear combination of $x_1, \dots, x_n$. The uniqueness requirement ensures that there are no two distinct $\mathbb{Z}$ -linear combinations in $FA(X)$ which map to the same element in F , and hence $\tilde{f}$ is injective as well. Therefore $F \cong FA(X)$, which is the free abelian group of rank n .

Conversely, suppose that $F \cong FA(X) \cong FA_n$ and that the elements of $X \subseteq F$ are $\mathbb{Z}$ -linearly independent.

Since X generates F we know straight away that any $g \in F$ can be written as a $\mathbb{Z}$ -linear combination

$$g = a_1x_1 + \cdots + a_nx_n$$

for some $a_1, \dots, a_n \in \mathbb{Z}$. Suppose there exists another $\mathbb{Z}$ -linear combination

$$g = b_1x_1 + \cdots + b_nx_n.$$

Subtracting one from the other yields

$$(a_1 - b_1)x_1 + \cdots + (a_n - b_n)x_n = 0$$

and since the generators $x_1, \dots, x_n$ are $\mathbb{Z}$ -linearly independent,

$$a_1 - b_1 = 0, \quad \dots, \quad a_n - b_n = 0$$

and therefore

$$a_1 = b_1, \quad \dots, \quad a_n = b_n.$$

Hence any such $\mathbb{Z}$ -linear combination for g must be unique. $\square$

Some pages ago we stated the Nielsen–Schreier Theorem,¹² which says that a subgroup of a free group is itself free. The proof of this is a little involved and is postponed until later in this chapter, but the corresponding result for free abelian groups is a little easier to prove:

¹² Theorem 5.11, page 143.

Theorem 5.46 (Richard Dedekind (1831–1916)) *Let $F \cong FA_r$ be a free abelian group of rank r , and let G be a subgroup of F . Then G is a free abelian group of rank at most r .*

Proof We will prove this theorem by induction on r . The base case ($r = 1$) yields $FA_r \cong \mathbb{Z}$ and we know that any subgroup of $\mathbb{Z}$ is either trivial or of the form $n\mathbb{Z} = \langle n \rangle$ for some $n \in \mathbb{Z}$. The trivial group can be regarded as $FA_0 = FA(\emptyset)$, the free abelian group with no generators. A cyclic subgroup $n\mathbb{Z} \leq \mathbb{Z}$ is itself isomorphic to $\mathbb{Z}$ by the mapping $nk \mapsto k$ for any element $nk \in n\mathbb{Z}$.

Now suppose that $r > 1$ and that the required statement holds for free abelian groups of rank $r-1$. Let $H = FA_{r-1}$ be the subgroup of $F = FA_r$ generated by the first $r-1$ generators $x_1, \dots, x_{r-1}$, and let $K \cong \mathbb{Z} \cong FA_1$ be the subgroup of F generated by the r th generator x_r :

$$H = \langle x_1, \dots, x_r \rangle \cong FA_{r-1} \quad \text{and} \quad K = \langle x_r \rangle \cong \mathbb{Z} \cong FA_1.$$

We can see that $F = H \times K$.

By Proposition 2.9 $G \cap H$ is a subgroup of H , and hence by the induction hypothesis it is also a free abelian group of rank $s \leq r-1$, with generators $y_1, \dots, y_s$. The Second Isomorphism Theorem¹³ tells us

$$G/(G \cap H) \cong (G + H)/H$$

where $G+H = \{g+h : g \in G, h \in H\}$.¹⁴

This set $G+H$ is a subgroup of F , since it contains the identity 0, it's closed under addition, and it contains all the required inverses. Hence $(G+H)/H$ is a subgroup of F/H . But F/H is exactly the subgroup $K = \langle x_r \rangle \cong FA_1$, and so we can apply the induction base case ($r = 1$) to see that $G/(G \cap H)$ is either trivial or isomorphic to $\mathbb{Z} \cong FA_1$.

If $G/(G \cap H)$ is trivial, then $G \cong G \cap H$ which we've already confirmed is free abelian with rank $s \leq r-1$.

Suppose, then, that $G/(G \cap H) \cong \mathbb{Z}$. This group consists of the cosets of the intersection $G \cap H$ in G , which are of the form $g+(G \cap H)$ for $g \in G$. Since $G/(G \cap H)$ is isomorphic to the infinite cyclic group $\mathbb{Z}$ there must exist some $g \in G \setminus H$ for which $G/(G \cap H)$ is generated by $g+(G \cap H)$; that is, $G/(G \cap H) \cong \langle g+(G \cap H) \rangle$.

Let $g = h + nx_r$ for some $h \in H$ and $n \in \mathbb{N}$. We claim that G is free abelian with basis $Y = \{y_1, \dots, y_s, g\}$. This set Y certainly generates G , but we need also to show that it's linearly independent over $\mathbb{Z}$. Therefore, suppose that

$$a_1y_1 + \dots + a_sy_s + ag = 0$$

where $a, a_1, \dots, a_s \in \mathbb{Z}$. Substituting in $g = h + nx_r$ gives

$$a_1y_1 + \dots + a_sy_s + ah + nax_r = 0 \quad (5.1)$$

and so

$$nax_r = -ah - (a_1y_1 + \dots + a_sy_s).$$

The left hand side of this equation is in $K = \langle x_r \rangle$, while the right hand side lies in H (since the $y_1, \dots, y_s \in G \cap H$ and $h \in H$). Therefore both sides of the equation must belong to both K and H , and hence to the intersection $H \cap K$. But by the definition of H and K we know that $H \cap K = \{0\}$, hence $nax_r = 0$. Since $n \neq 0$, the only other possibility is that $a = 0$. Substituting this back into (5.1) we get

$$a_1y_1 + \dots + a_sy_s = 0.$$

The $y_1, \dots, y_s$ are free abelian generators of the intersection $G \cap H$, and hence by the $\mathbb{Z}$ -linear independence condition (Proposition 5.45) it must be the case that $a_1 = \dots = a_s = 0$. This means that every element of G can be uniquely expressed as a $\mathbb{Z}$ -linear combination of the elements $y_1, \dots, y_s, g \in Y$, and hence $G \cong FA(Y)$ and $\text{rank}(G) = s + 1 \leq r$. $\square$

¹³ Theorem 4.62, page 126.

¹⁴ Definition 2.10, page 45.

Returning to our discussion of presentations of direct products and direct sums, obviously if both groups G and H are abelian, then so is their direct sum $G \oplus H$. We can see this from the presentation for $G \oplus H$ given in Proposition 5.22. A presentation of the form

$$\begin{aligned} \mathbb{Z}_{n_1} \oplus \cdots \mathbb{Z}_{n_k} \oplus \mathbb{Z} \oplus \cdots \oplus \mathbb{Z} = \\ \langle x_1 : x_1^{n_1} \rangle \oplus \cdots \oplus \langle x_k : x_k^{n_k} \rangle \oplus \langle x_{k+1} : \rangle \oplus \cdots \oplus \langle x_m : \rangle = \\ \langle x_1, \dots, x_m : x_1^{n_1} = \cdots = x_k^{n_k} = e, [x_i, x_j] = e \text{ for } 1 \leq i, j \leq m \rangle \end{aligned}$$

gives an abelian group. Less obviously, *every* abelian group with a finite presentation is of this form: all finitely-generated abelian groups are isomorphic to a direct sum of a finite number of cyclic groups.

The proof of this classification is a little involved, and we will split it up into a few more manageable pieces. First we introduce a few terms that will be used in the statement of the theorem and the following discussion and proof.

Definition 5.47 Suppose that

$$G \cong \mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_k} \oplus \mathbb{Z}^r$$

where $m_1, \dots, m_k, r \in \mathbb{Z}, r \geq 0$ and $m_i | m_{i+1}$ for $1 \leq i < k$.

This is called an **invariant factor decomposition** of G , with **torsion coefficients** or **invariant factors** $m_1, \dots, m_k$ and **rank** r .

The statement of the classification theorem is as follows:

Theorem 5.48 *Let G be a finitely-generated abelian group. Then G has a unique invariant factor decomposition*

$$G \cong \mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_k} \oplus \mathbb{Z}^r$$

in the sense that if H is another finitely-generated abelian group with invariant factor decomposition

$$H \cong \mathbb{Z}_{n_1} \oplus \cdots \oplus \mathbb{Z}_{n_l} \oplus \mathbb{Z}^s,$$

then $G \cong H$ if and only if $k = l, r = s$ and $m_i = n_i$ for $1 \leq i < k$.

Two important special cases are given by the following corollaries:

Corollary 5.49 *Let G be a finite abelian group. Then G has a unique invariant factor decomposition*

$$G \cong \mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_k}.$$

Corollary 5.50 *Let G be a torsion-free abelian group. Then $G \cong \mathbb{Z}^r$ for some non-negative integer r .*

Before we get to the proof of this theorem, we'll look at a few examples to see how it all works.

Example 5.51 Let G be an abelian group of order 12. Then since G must be a direct sum of finite cyclic groups, we have the following possibilities:

$$\mathbb{Z}_{12}, \quad \mathbb{Z}_2 \oplus \mathbb{Z}_6, \quad \mathbb{Z}_3 \oplus \mathbb{Z}_4, \quad \text{and} \quad \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3.$$

But by Proposition 1.32 we know that

$$\mathbb{Z}_3 \oplus \mathbb{Z}_4 \cong \mathbb{Z}_{12} \quad \text{and} \quad \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_6.$$

Only two of these four satisfy the invariant factor condition, namely $\mathbb{Z}_{12}$ and $\mathbb{Z}_2 \oplus \mathbb{Z}_6$.

Example 5.52 Let's list all the abelian groups of order 48. The integer 48 factorises as $2 \times 2 \times 2 \times 2 \times 3$, so we have the following possibilities:

$$\begin{array}{lll} \mathbb{Z}_{48}, & \mathbb{Z}_2 \oplus \mathbb{Z}_{24}, & \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_{12} \\ \mathbb{Z}_4 \oplus \mathbb{Z}_{12}, & \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_6, & \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6, \\ \mathbb{Z}_8 \oplus \mathbb{Z}_6, & \mathbb{Z}_8 \oplus \mathbb{Z}_6, & \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3, \\ \mathbb{Z}_{16} \oplus \mathbb{Z}_3, & \mathbb{Z}_2 \oplus \mathbb{Z}_8 \oplus \mathbb{Z}_3, & \text{and} \quad \mathbb{Z}_4 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_3. \end{array}$$

From Proposition 1.32 we know that

$$\begin{aligned} \mathbb{Z}_{48} &\cong \mathbb{Z}_{16} \oplus \mathbb{Z}_3, \\ \mathbb{Z}_2 \oplus \mathbb{Z}_{24} &\cong \mathbb{Z}_8 \oplus \mathbb{Z}_6 \\ &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_8 \oplus \mathbb{Z}_3, \\ \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_{12} &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_6 \\ &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_3, \\ \mathbb{Z}_4 \oplus \mathbb{Z}_{12} &\cong \mathbb{Z}_4 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_3, \\ \text{and} \quad \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6 &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3. \end{aligned}$$

There are therefore five different abelian groups of order 48, whose invariant factor decompositions are as follows:

$$\mathbb{Z}_{48}, \quad \mathbb{Z}_2 \oplus \mathbb{Z}_{24}, \quad \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_{12}, \quad \mathbb{Z}_4 \oplus \mathbb{Z}_{12} \quad \text{and} \quad \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6.$$

We'll split the proof of Theorem 5.48 into two parts: existence (every finitely generated abelian group G has an invariant factor decomposition) and uniqueness (isomorphic finitely generated abelian groups have the same invariant factor decomposition).

In order to prove the first part, we'll need the following lemma, which is really just a consequence of Proposition 5.35 in an abelian setting.

Lemma 5.53 Let $X = \{x_1, \dots, x_n\}$, and let

$$Y = \{x_1, \dots, x_{j-1}, x_j + mx_i, x_{j+1}, \dots, x_n\}$$

where $1 \leq i \neq j \leq n$ and $m \in \mathbb{Z}$.

Then if X is a basis for some abelian group G , so is Y .

So, for an abelian group G we can replace any generator x_j with a new generator of the form $x_j + mx_i$ where $i \neq j$ and $m \in \mathbb{Z}$.

Proof This is just an application of Tietze transformations. Replacing the generator x_j with a new generator $y = x_j + mx_i$ amounts to an X_+ move introducing a generator y and a relator $y - x_j - mx_i$, followed by an X_- move removing the generator x_j and the relator $y - x_j - mx_i$. $\square$

Proof of Theorem 5.48 (existence) Our abelian group G has generators $x_1, \dots, x_s$, and relators $r_1, \dots, r_t$, each of which are of the form

$$r_i = n_{i,1}x_1 + \dots + n_{i,s}x_s.$$

If $t = 0$ (that is, if there are no relators) then we have the free abelian group FA_s of rank s , which is isomorphic to $\mathbb{Z}^s$ by Proposition 5.44.

Suppose instead that $t \geq 1$. Then there exists at least one nontrivial relator, and amongst those relators there will be a smallest positive coefficient $n_{i,j}$ attached to the generator x_j in relator r_i .

We can relabel the generators and relators¹⁵ so that this coefficient is $n_{1,1}$ and we're working with the relator

$$r_1 = n_{1,1}x_1 + \dots + n_{1,s}x_s.$$

¹⁵ This won't change the group under investigation, except by an isomorphism, the details of which don't matter to this discussion.

Call this coefficient m_1 , so our relator becomes

$$m_1x_1 + n_{1,2}x_2 + \dots + n_{1,s}x_s.$$

This m_1 is intended to be the first torsion coefficient in the statement of the theorem, and we want to show it's a factor of $n_{1,2}$.

Rewrite $n_{1,2} = am_1 + b$, where $0 \leq b < m_1$. Then our relator becomes

$$\begin{aligned} m_1x_1 + (am_1 + b)x_2 + n_{1,3}x_3 + \dots + n_{1,s}x_s \\ = m_1x_1 + a(m_1x_2) + bx_2 + n_{1,3}x_3 + \dots + n_{1,s}x_s \\ = m_1(x_1 + ax_2) + bx_2 + n_{1,3}x_3 + \dots + n_{1,s}x_s. \end{aligned}$$

Lemma 5.53 tells us that $(x_1 + ax_2), x_2, \dots, x_s$ is also a minimal set of generators for G . But $b < m_1$, contradicting our statement that m_1 is the smallest positive coefficient over all generators and relators. Thus $b = 0$, so $n_{1,2} = am_1$, and hence $m_1 | n_{1,2}$. In a similar way, we can show that m_1 divides each of the other coefficients $n_{1,3}, \dots, n_{1,s}$, and so in each case we have $n_{1,j} = a_j m_1$ where $a_j \in \mathbb{Z}$ and $2 \leq j \leq s$.

We can now perform Tietze transformations to replace the generator x_1 with a new generator

$$z_1 = x_1 + a_2x_2 + a_3x_3 + \dots + a_sx_s$$

and rewrite the relation

$$r_1 = m_1x_1 + n_{1,2}x_2 + \dots + n_{1,s}x_s = 0$$

as $mz_1 = 0$.

Because of our choice of m_1 , no smaller multiple of z_1 is zero, so z_1 has order m_1 .

Now let $G_1 = \langle z_1 \rangle \cong \mathbb{Z}_{m_1}$ be the subgroup of G generated by z_1 , and let H be the subgroup of G generated by the other generators $x_2, \dots, x_s$. We can see that $G_1 + H = G$ and that $G_1 \cap H = \{0\}$, so Proposition 2.12 tells us that $G \cong G_1 \oplus H$.

We can now apply the same procedure to H , with two possible outcomes. One possibility is that H is a free abelian group of rank $s-1$, and hence isomorphic to $\mathbb{Z}^{s-1}$, in which case $G \cong \mathbb{Z}_{m_1} \oplus \mathbb{Z}^{s-1}$. If so, then we've finished: G is isomorphic to a group of the required form.

The other possibility is that $H \cong \mathbb{Z}_{m_2} \oplus K$ where K is generated by $s-2$ generators, and in this case we have to do a little bit more work, to show that m_2 is a multiple of m_1 .

So, suppose that m_2 is the smallest positive coefficient in all of the relators $r_2, \dots, r_t$ in $H = \langle x_2, \dots, x_s : r_2, \dots, r_t \rangle$, and that (possibly after some reordering of generators and relators) it happens to be the coefficient of the generator x_2 in the relator

$$r_2 = m_2 x_2 + n_{2,3} x_3 + \cdots + n_{2,s} x_s.$$

Since $z_1, x_2, \dots, x_s$ is a minimal set of generators for G , and since

$$m_1 z_1 + m_2 x_2 + n_{2,3} x_3 + \cdots + n_{2,s} x_s = 0$$

we know m_1 has to be a factor of m_2 by the argument we used earlier.

At this point we have all we need to complete the proof. We know that m_1 is a factor of m_2 , and by applying the same procedure to the group H we'll eventually end up with the direct sum

$$G \cong \mathbb{Z}_{m_1} \oplus \mathbb{Z}_{m_2} \oplus \cdots \oplus \mathbb{Z}_{m_k} \oplus \mathbb{Z}^r$$

where $m_1 | m_2 | \cdots | m_k$ as required. $\square$

So, we now know that any finitely-generated abelian group can be decomposed as a direct sum in the specified way. Now we need to show that this decomposition is unique. First we need a short lemma.

Lemma 5.54 *There are $\gcd(n, q)$ elements $k \in \mathbb{Z}_n$ for which $qk = 0$. If $G = \mathbb{Z}_{n_1} \oplus \cdots \oplus \mathbb{Z}_{n_r}$, then there are $\gcd(n_1, q) \times \cdots \times \gcd(n_r, q)$ elements $k \in G$ for which $qk = 0$.*

Proof If $qk = 0$ then this is equivalent to saying that $n | qk$. Suppose $d = \gcd(n, q)$ and let $n = md$ and $q = pd$, so that $\gcd(m, p) = 1$. If $n | qk$ then $md | pdk$ and so $m | k$. Then k must be one of $0, m, 2m, \dots, (d-1)m$. There are $d = \gcd(n, q)$ of these, which proves the first part.

For the second part, observe that an arbitrary element of the group $G = \mathbb{Z}_{n_1} \oplus \cdots \oplus \mathbb{Z}_{n_r}$ has the form $k = (k_1, \dots, k_r)$. Then if $qk = q(k_1, \dots, k_r) = (qk_1, \dots, qk_r) = (0, \dots, 0)$, we have $\gcd(n_1, q)$ valid

choices for k_1 , together with $\gcd(n_2, q)$ valid choices for k_2 , all the way up to k_r , for which there are $\gcd(n_r, q)$ possibilities. Hence the number of elements $k = (k_1, \dots, k_r)$ satisfying $qk = 0$ is equal to the product of the greatest common divisors, namely $\gcd(n_1, q) \times \dots \times \gcd(n_r, q)$. $\square$

Proof of Theorem 5.48 (uniqueness) We will prove uniqueness in two parts, first by considering the infinite-order component, and then by considering the torsion coefficients $m_1, \dots, m_k$.

Let $G_{\text{fin}} = \mathbb{Z}_{m_1} \oplus \dots \oplus \mathbb{Z}_{m_k}$ and $G_{\text{inf}} = \mathbb{Z}^r$, and let $H_{\text{fin}} = \mathbb{Z}_{n_1} \oplus \dots \oplus \mathbb{Z}_{n_l}$ and $H_{\text{inf}} = \mathbb{Z}^s$, so that $G = G_{\text{fin}} \oplus G_{\text{inf}}$ and $H = H_{\text{fin}} \oplus H_{\text{inf}}$.

There is a surjective projection homomorphism $f_1: G \rightarrow G_{\text{inf}}$ which maps $(g_1, \dots, g_k, g_{k+1}, \dots, g_r) \mapsto (g_{k+1}, \dots, g_r)$. The kernel $\ker(f_1)$ of this homomorphism is obviously the finite-order component G_{fin} , and its image is equally obviously the infinite-order component $G_{\text{inf}} = \mathbb{Z}^r$.

By the First Isomorphism Theorem¹⁶ we see that

¹⁶ Theorem 4.40, page 117.

$$G / \ker(f_1) = G / G_{\text{fin}} \cong \text{im}(f_1) = G_{\text{inf}} = \mathbb{Z}^r.$$

We can define a similar homomorphism $f_2: H \rightarrow H_{\text{inf}}$ which maps $(h_1, \dots, h_l, h_{l+1}, \dots, h_s) \mapsto (h_{l+1}, \dots, h_s)$, and use this to show that

$$H / \ker(f_2) = H / H_{\text{fin}} \cong \text{im}(f_2) = H_{\text{inf}} = \mathbb{Z}^s.$$

Any isomorphism $\phi: G \rightarrow H$ must map the elements of finite order in G to the elements of finite order in H , by Proposition 1.21, so $G_{\text{fin}} \cong H_{\text{fin}}$, and hence there must be an induced isomorphism

$$\phi_*: G / G_{\text{fin}} = G_{\text{inf}} \rightarrow H / H_{\text{fin}} = H_{\text{inf}},$$

which means that $\mathbb{Z}^r \cong \mathbb{Z}^s$ and so $r = s$ as claimed.

All that remains is to prove uniqueness of the torsion coefficients. As just noted, $G_{\text{fin}} \cong H_{\text{fin}}$. Suppose, without loss of generality, that $k \geq l$. Then we apply Lemma 5.54 to G_{fin} and H_{fin} with $q = m_1$ to see that

$$\gcd(m_1, m_1) \times \dots \times \gcd(m_1, m_k) = \gcd(m_1, n_1) \times \dots \times \gcd(m_1, n_l)$$

and since $m_1 | m_2 | \dots | m_k$ we have

$$\gcd(m_1, m_1) = \gcd(m_1, m_2) = \dots = \gcd(m_1, m_k) = m_1$$

and therefore

$$m_1^k = \gcd(m_1, n_1) \times \dots \times \gcd(m_1, n_l).$$

Each of the factors $\gcd(m_1, n_1), \dots, \gcd(m_1, n_l)$ on the right hand side can be at most m_1 , which forces $k = l$. This in turn means that $\gcd(m_1, n_1) = m_1$, and hence m_1 must be a factor of n_1 .

Now apply Lemma 5.54 with $q = n_1$ to get

$$\gcd(m_1, n_1) \times \dots \times \gcd(m_k, n_1) = n_1^k$$

so $\gcd(m_1, n_1) = n_1$ and thus n_1 divides m_1 . Therefore $m_1 = n_1$.

Next we can apply Lemma 5.54 again with $q = m_2$ to get

$$m_1 m_2^{k-1} = \gcd(m_2, n_1) \times \cdots \times \gcd(m_2, n_k)$$

which implies

$$m_2^{k-1} = \gcd(m_2, n_2) \times \cdots \times \gcd(m_2, n_k)$$

from which we see that $\gcd(m_2, n_2) = m_2$, so $m_2 | n_2$.

Applying Lemma 5.54 yet again with $q = n_2$ we get

$$m_1 \times \gcd(m_2, n_2) \times \cdots \times \gcd(m_k, n_2) = m_1 n_2^{k-1}$$

which tells us that $n_2 | m_2$, hence $m_2 = n_2$.

Repeating this process another $k-2$ times shows that

$$n_1 = m_1, \dots, n_k = m_k,$$

which completes the proof. $\square$

This classification theorem is elegant and useful. It tells us that every finitely generated abelian group has a unique canonical form as a direct sum of cyclic groups. What we now want is a general procedure for finding that decomposition from a given presentation.

The key observation is that the problem reduces to solving a finite system of linear simultaneous equations with integer coefficients. Basic linear algebra gives us techniques for solving systems of linear equations with real coefficients, by transforming an augmented matrix into reduced row echelon form by a finite sequence of elementary row operations. We have to be a little careful working with integer coefficients, but the procedure is similar: we convert the relations into a suitable matrix, use elementary row operations to convert that matrix into **Smith normal form**, and then read off the torsion coefficients.

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & a_{rr} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Figure 5.1: A matrix in Smith normal form

Definition 5.55 An $m \times n$ matrix A of rank r is in **Smith normal form** if every element is zero, except possibly for the diagonal elements a_{ii} for $1 \leq i \leq r$, and further that $a_{ii} | a_{i+1, i+1}$ for $1 \leq i < r$.

That is, A is of the form shown in Figure 5.1, where the last $n-r$ columns and the last $m-r$ rows are all zero.

Example 5.56 The following matrices are in Smith normal form:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 12 \\ 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 7 & 0 & 0 & 0 \\ 0 & 14 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 4 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

When solving systems of simultaneous equations with real coefficients, we employ elementary row and column operations to convert an augmented matrix to row echelon form, and the same principle applies here. First we need to define the appropriate elementary operations:

Definition 5.57 Let A be an $m \times n$ matrix with integer entries. Then we may apply one or more of the following **elementary row and column operations** to A in order to obtain a similar matrix.

- E1** Swap two rows or two columns.
- E2** Multiply all entries of a row or column by -1 .
- E3** Add an integer multiple of one row to another row, or an integer multiple of one column to another column.

The next proposition is the key to the whole problem.

Proposition 5.58 Any $m \times n$ integer matrix A can be transformed into Smith normal form by a finite sequence of row and column operations of type E_1 , E_2 and E_3 .

Equivalently, there exists an $m \times m$ matrix P and an $n \times n$ matrix Q such that PAQ is an $m \times n$ matrix in Smith normal form.

The following algorithm provides a constructive proof of this fact.

Algorithm 5.59 Assume that A is not already in Smith normal form.

- 1 Use row and column operations of type E_1 to arrange the matrix so that the nonzero entries in the first row and first column are in ascending order of absolute value, followed by any zero elements, so that $|a_{11}| \leq |a_{12}| \leq \dots \leq |a_{1s}|$ and $|a_{11}| \leq |a_{21}| \leq \dots \leq |a_{t1}|$.
- 2 Use row and column operations of type E_2 to ensure that all of the entries $a_{11}, \dots, a_{1s}$ in the first row are positive, and all of the entries $a_{11}, \dots, a_{t1}$ in the first column are also positive.
- 3 If a_{11} divides all other nonzero entries in row 1, then go to step 4. Otherwise, let a_{1j} be the first nonzero entry in the first row which isn't an integer multiple of a_{11} . Then we can find non-negative integers q and p such that $a_{1j} = qa_{11} + p$ where $0 \leq p < a_{11}$. Apply a column operation of type E_3 , subtracting q times column 1 from column j . Repeat this process for all the other columns for which the first element is not an integer multiple of a_{11} . Go to step 1.
- 4 If a_{11} divides all of the other nonzero entries in the first column, then go to step 5. Otherwise, let a_{k1} be the first nonzero entry in column 1 that isn't an integer multiple of a_{11} . Then we can find non-negative integers q and p such that $a_{k1} = qa_{11} + p$ where $0 \leq p < a_{11}$. Apply an E_3 row operation, subtracting q times row 1 from row k . Repeat this process for all the other rows for which the first element is not an integer multiple of a_{11} . Go to step 1.



Wikimedia Commons

Henry John Stephen Smith (1826–1883) was born in Dublin in late 1826, the fourth child of a barrister named John Smith. After his father's death in 1829 the family moved to England, where he was educated by private tutors until the age of fifteen. In 1841 he was admitted to Rugby School in Warwickshire, a prestigious English public (which is to say, private) school founded in 1567. At the time, the school's headmaster was the renowned theologian and educator Dr Thomas Arnold (1795–1842). A fictionalised portrait of Dr Arnold and the school can be found in the classic 1857 novel *Tom Brown's School Days* by Smith's fellow pupil Thomas Hughes (1822–1895). In 1845, Smith won a scholarship to Balliol College, Oxford, where he studied mathematics and classics, graduating with high honours in 1849 and being promoted to Fellow shortly afterwards. In 1861 he was appointed Savilian Professor of Geometry and elected a Fellow of both the Royal Society of London and the Royal Astronomical Society. Until his death in 1883, he taught and pursued research in mathematics, making many important contributions to geometry and number theory. Smith was well-respected by his colleagues. John Conington (1825–1869), Professor of Latin, once remarked: "I do not know what Henry Smith may be at the subjects of which he professes to know something; but I never go to him about a matter of scholarship, in a line where he professes to know nothing, without learning more from him than I can get from any one else."

- 5 Every entry in row 1 and column 1 is now a multiple of a_{11} .
 Now apply column operations of type E₃, subtracting multiples of column 1 from each of the other columns with nonzero first element, so that a_{11} is the only nonzero element on the first row. Similarly, apply row operations of type E₃, subtracting multiples of row 1 from all the other rows that have a nonzero first element, so that a_{11} is left as the only nonzero element in column 1.

The matrix is then in the form

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

and we can apply steps 1–5 to the $(m-1) \times (n-1)$ submatrix

$$\begin{bmatrix} a_{22} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

to get an $m \times n$ matrix where the only nonzero elements in the first two rows and columns are on the diagonal. We repeat this until we get a matrix whose only nonzero entries are on the diagonal. These diagonal entries won't necessarily satisfy the divisibility criterion $a_{11} | a_{22} | \dots | a_{rr}$. If they don't, we proceed to the next step.

- 6 For $1 \leq i \leq r-1$, compare a_{ii} with each a_{jj} for $i < j \leq r$. Let i and j be the lowest integers for which a_{ii} doesn't divide a_{jj} . Use a row operation of type E₃ to add row j to row i , and then reduce this new $m \times n$ matrix using steps 1–5.

This algorithm will eventually terminate, yielding a matrix in Smith normal form.¹⁷ Let's try a couple of examples.

¹⁷ A slightly more efficient step 6 is described in a paper by VJ Rayward-Smith,¹⁸ but this will suffice for our purposes.

¹⁸ VJ Rayward-Smith, *On computing the Smith normal form of an integer matrix*, ACM Transactions on Mathematical Software 5.4 (1979) 451–456

Example 5.60 Let $\mathbb{Z}_2 \oplus \mathbb{Z}_3$ be the abelian group with generators x and y , and relations

$$2x = 0, \quad 3y = 0.$$

This yields the coefficient matrix

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}.$$

Applying Algorithm 5.59, we see A is already diagonal, so we skip to step 6. Now $a_{11} = 2$, which doesn't divide $a_{22} = 3$, so we perform an operation of type E₃, adding row 2 to row 1, to get the matrix

$$\begin{bmatrix} 2 & 3 \\ 0 & 3 \end{bmatrix}.$$

Starting again from step 1, the matrix evolves as follows:

$$\begin{bmatrix} 2 & 3 \\ 0 & 3 \end{bmatrix} \mapsto \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 \\ 3 & -6 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 \\ 0 & -6 \end{bmatrix}.$$

We can perform a final E_2 move to get

$$\begin{bmatrix} 1 & 0 \\ 0 & 6 \end{bmatrix},$$

and then read off the new relations

$$x = 0, \quad 6y = 0$$

which yields the group $\mathbb{Z}_6$, as expected from Proposition 1.32.

Now let's try a more complicated example.

Example 5.61 Let G be the abelian group with four generators w, x, y and z , and three relations

$$16w + 56x + 4y + 48z = 0,$$

$$4w + 16x + 4y - 8z = 0,$$

$$10w + 22x - 2y + 70z = 0.$$

Under the application of Algorithm 5.59, the coefficient matrix evolves as follows:

$$\begin{aligned} & \begin{bmatrix} 16 & 56 & 4 & 48 \\ 4 & 16 & 4 & -8 \\ 10 & 22 & -2 & 70 \end{bmatrix} \xrightarrow{E_1} \begin{bmatrix} -2 & 10 & 22 & 70 \\ 4 & 4 & 16 & -8 \\ 4 & 16 & 56 & 48 \end{bmatrix} \xrightarrow{E_2} \begin{bmatrix} 2 & 10 & 22 & 70 \\ 4 & -4 & -16 & 8 \\ 4 & -16 & -56 & -48 \end{bmatrix} \\ & \xrightarrow{E_3} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 4 & -24 & -60 & -132 \\ 4 & -36 & -100 & -188 \end{bmatrix} \xrightarrow{E_2} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 24 & 60 & 132 \\ 0 & 36 & 100 & 188 \end{bmatrix} \xrightarrow{E_3} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 24 & 12 & 12 \\ 0 & 36 & 28 & 8 \end{bmatrix} \\ & \xrightarrow{E_1} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 8 & 28 & 36 \\ 0 & 12 & 12 & 24 \end{bmatrix} \xrightarrow{E_3} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 8 & 4 & 4 \\ 0 & 12 & -24 & -24 \end{bmatrix} \xrightarrow{E_1} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 4 & 8 \\ 0 & -24 & -24 & 12 \end{bmatrix} \\ & \xrightarrow{E_2} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 4 & 8 \\ 0 & 24 & 24 & -12 \end{bmatrix} \xrightarrow{E_3} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 24 & 0 & -60 \end{bmatrix} \xrightarrow{E_1} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 24 & -60 & 0 \end{bmatrix} \\ & \xrightarrow{E_3} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & -60 & 0 \end{bmatrix} \xrightarrow{E_2} \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 60 & 0 \end{bmatrix} \end{aligned}$$

The corresponding relations are therefore

$$2w = 0, \quad 4x = 0, \quad 60y = 0.$$

The fourth generator z is free, since it has no associated relation, and hence $G \cong \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_{60} \oplus \mathbb{Z}$.

There is another classification theorem for finitely-generated abelian

groups, in which the torsion subgroups are grouped in a different way. We'll state the theorem, but leave the proof as an exercise.

Definition 5.62 Suppose that

$$G \cong \mathbb{Z}_{p_1^{n_1}} \oplus \cdots \oplus \mathbb{Z}_{p_k^{n_k}} \oplus \mathbb{Z}^r$$

where $p_1, \dots, p_k \in \mathbb{N}$ are (not necessarily distinct) prime integers, and $n_1, \dots, n_k, r \in \mathbb{Z}$.

This is called a **primary decomposition** of G with **torsion coefficients** $p_1^{n_1}, \dots, p_k^{n_k}$ and **rank** r .

Theorem 5.63 A finitely-generated abelian group G has a unique primary decomposition

$$G \cong \mathbb{Z}_{p_1^{n_1}} \oplus \cdots \oplus \mathbb{Z}_{p_k^{n_k}} \oplus \mathbb{Z}^r$$

in the sense that if H is another finitely-generated abelian group with primary decomposition

$$H \cong \mathbb{Z}_{q_1^{m_1}} \oplus \cdots \oplus \mathbb{Z}_{q_l^{m_l}} \oplus \mathbb{Z}^s$$

then $G \cong H$ if and only if $k = l$, $r = s$ and the primes $p_1, \dots, p_k$ are equal to the primes $q_1, \dots, q_l$, up to possible rearrangement.

Unfortunately there isn't a neat classification of finitely-generated nonabelian groups, although there are many partial results and we will see a few of them later, including the Jordan–Hölder Theorem on composition series, which will be covered in Chapter 7.

Count what is countable, measure what is measurable, and what is not measurable, make measurable.

— attributed to Galileo Galilei
(1564–1642)

5.A Coset enumeration

PRESENTATIONS provide a neat and compact way of describing many classes of groups, both finite and infinite, but there are technical difficulties that sometimes limit their usefulness. We can't always deduce everything we might want to know about a given group G : indeed it's not always possible to calculate the order of G , or even say whether it's finite or infinite. In many cases we can obtain an upper bound on $|G|$ by using the technique illustrated in Proposition 5.32.

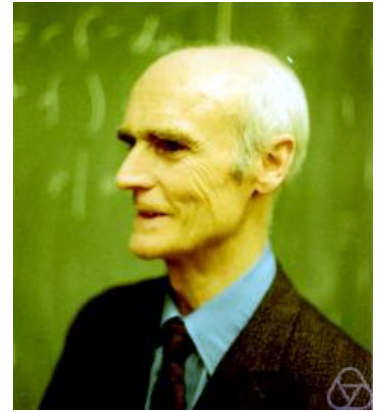
Example 5.64 Let $G = \langle x, y : x^2, y^5, (xy)^2 \rangle$. Then the relations $x^2 = e$ and $y^5 = e$ tell us that $x^{-1} = x$ and $y^{-1} = y^4$, so we need only consider non-negative powers of x and y .

The relation $(xy)^2 = e$ means that $xy = (xy)^{-1} = y^{-1}x^{-1} = y^4x$. So we can rewrite any string of x s and y s so as to group the x and y terms together. Hence any element of G is of the form $x^m y^n$ where

$0 \leq m \leq 1$ and $0 \leq y \leq 4$. There are ten possibilities, so $|G| \leq 10$.

(As it happens, we can see from examining the presentation that $G \cong D_5$ so we know that actually $|G| = 10$.)

There are more sophisticated methods for deriving structural information from group presentations, and next we'll look at one of them, the **Todd–Coxeter Algorithm**. This algorithm was discovered by J A Todd (1908–1994) and HSM Coxeter (1907–2003) in 1936, and in some cases can provide structural information about a finitely presented group. We won't formally state the algorithm, instead we'll illustrate it with worked examples. We start with a simple case: the cyclic group $\mathbb{Z}_3$.



Oberwolfach Photo Collection / Konrad Jacobs

The British mathematician **Harold Scott MacDonal Coxeter** (1907–2003) was born in London and studied at Cambridge, where he was awarded a PhD in 1931 for a thesis entitled *Some Contributions to the Theory of Regular Polytopes*. In 1936 he took up a post at the University of Toronto, where he remained for the rest of his long career. Known as Donald to friends, family and colleagues, he developed a keen interest in both music and mathematics and by the age of twelve he was an accomplished pianist and amateur composer.

His main contributions were in the field of geometry and algebra. In particular, since his schooldays he had been interested in polyhedra and their higher-dimensional analogues, **polytopes**, subjects which he often discussed with Alicia Boole Stott (1860–1940), a family friend and the daughter of the mathematicians George Boole (1815–1864) and Mary Everest Boole (1832–1916).

Discussions with the Dutch artist Maurits Cornelis Escher (1898–1972) inspired some of the latter's work, in particular the *Circle Limit* series of woodcuts, based on tessellations of the hyperbolic plane.

He was elected to the Royal Society of Canada in 1950, the Royal Society of London in 1950, the American Academy of Arts and Sciences in 1990, and appointed a Companion of the Order of Canada in 1997.

He attributed his long and active life to his vegetarian diet, and a regime of regular exercise which he followed into his nineties.

Example 5.65 Let $G = \langle x : x^3 \rangle \cong \mathbb{Z}_3$. We expand the relator x^3 as a string of generators and inverses, obtaining xxx . We now construct a **relator table** with this string as the header, and with a 1 at the beginning and end of the first row:

x	x	x
1		1

Next pick an empty space next to a 1, write a 2 in it, draw a minus sign – (denoting that this is a definition) across the dividing line between the 1 and the 2 and add another row with a 2 at each end.

x	x	x
1	2	1
2		2

At this point we also start compiling a **coset table**:

definition	deduction
$1x = 2$	

The expression $1x = 2$ describes how the symbols 1 and 2 are related: it says that 2 (which represents x) equals 1 (which represents the identity e) multiplied on the right by x .

Returning to the relator table, we write a 2 in any empty cell to the right of a 1, and a 1 in any empty cell to the left of a 2 where the dividing line is labeled x . We also (although this is not applicable here) write a 2 in any empty cell to the left of a 1, and a 1 in any empty cell to the right of a 2 where the dividing line is labeled x^{-1} .

x	x	x	definition	deduction
1	2	1		$1x = 2$
2		2		

Now we write the symbol 3 in an available empty space in the relator table, add another row with a 3 in the first and last place, write the definition $2x = 3$ in the left-hand column of the coset table, and then insert 3s in appropriate places in the relator table:

x	x	x		definition	deduction
1	2	3	1		$1x = 2$
2	3	1	2		$2x = 3$
3			3		

At this point we see that the first row is full. In particular, this yields the expression $3x = 1$ for free: we write this in the deduction column of the coset table, and mark it in the relator table by an equals sign $=$ across the appropriate dividing line:

x	x	x		definition	deduction
1	2	3	1		$1x = 2$
2	3	1	2		$2x = 3$
3			3		$3x = 1$

Now we go through the relator table again, filling in any empty spaces with appropriate symbols. If we get any extra information then we mark this with an equals sign $=$ across the vertical line and list the new expression in the deduction column of the coset table. If a line completes without yielding any new information then we mark this with an equivalence sign $\equiv$.

x	x	x		definition	deduction
1	2	3	1		$1x = 2$
2	3	1	2		$2x = 3$
3	1	2	3		$3x = 1$

We can derive some useful information from the tables produced by this process (which is called **scanning**).

The number of rows in the completed relator table gives the order of the group G : in this case there are three rows, so $|G| = 3$ as expected.

We can also use the first two columns of the relator table to obtain a permutation in S_3 representing the generator x . In this case, we see that $x \mapsto (123) \in S_3$.

The data obtained by this algorithm also enables us to construct a diagram showing how the group generators interact with each other.

Definition 5.66 A **Cayley graph** or **Cayley diagram** for a group G is a directed graph which has one vertex for each element of the group. Whenever $h = g * x$ for two elements $g, h \in G$ and some

generator x , we draw a directed edge from the vertex g to the vertex h and label it with the generator x .

In particular, each relator can be interpreted as a closed loop in the Cayley graph, based at the vertex representing the identity element.

The relator tables obtained via the Todd–Coxeter algorithm enable us to draw a Cayley graph in a straightforward way. Each numeric symbol in the tables yields a vertex, and by reading along the rows of the relator tables we can see where to draw the directed edges.

For the group $G = \langle x : x^3 \rangle \cong \mathbb{Z}_3$ studied in Example 5.65 we get the diagram shown in Figure 5.2.

This procedure extends to presentations with more than one relator: we have a relator and coset table for each relator, and must keep all tables up to date at each step. The following example illustrates this.

Example 5.67 Let $G = \langle x, y : x^2, y^2, xyx^{-1}y^{-1} \rangle$. This is isomorphic to the Klein group $V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

$x \quad x$		$y \quad y$		$x \quad y \quad x^{-1} \quad y^{-1}$			
1	1	1	1	1			1

We begin with the first relator, writing a 2 in the available space in the relator table, and adding the definition $1x = 2$ to the first column in the coset table. This yields the deduction $2x = 1$. We then add a new row to the other relator tables and fill in 1s and 2s as appropriate.

$x \quad x$		$y \quad y$		$x \quad y \quad x^{-1} \quad y^{-1}$			
1	2	1	1	1	2		1
2	1	2	2	2			2
definition				deduction			
$1x = 2$				$2x = 1$			

Next, add a 3 to the first empty space in the second table, add a new definition $1y = 3$, and fill in the relator tables as far as possible.

$x \quad x$		$y \quad y$		$x \quad y \quad x^{-1} \quad y^{-1}$			
1	2	1	3	1	2		3
2	1	2	2	2			2
3		3	1	3		1	3
definition				deduction			
$1x = 2$				$2x = 1$			
$1y = 3$				$3y = 1$			

Now we write a 4 in the empty cell in the second row of the second

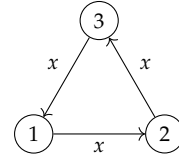


Figure 5.2: Cayley graph for the group $\langle x : x^3 \rangle \cong \mathbb{Z}_3$, constructed from the relator tables produced by the Todd–Coxeter algorithm. Here the node 1 corresponds to the identity element e , the node 2 corresponds to the element x , and the node 3 corresponds to the element x^2 .

table, and append the definition $2y = 4$ to the coset table. After filling in the relator tables as far as possible, we have the following:

x	x	y	y	x	y	x^{-1}	y^{-1}
1	2	1	1	1	2	4	3
2	1	2	2	2	1	3	4
3	4	3	3	3	4	2	1
4	3	4	4	4	3	1	2

definition deduction

$1x = 2$	$2x = 1$
$1y = 3$	$3y = 1$
$2y = 4$	$4y = 2$
	$3x = 4$
	$4x = 3$

The relator tables are now complete, and $|G| = 4$ as expected. We also get permutation representatives in S_4 for the generators of G :

$$x \mapsto (12)(34) \quad \text{and} \quad y \mapsto (13)(24)$$

Figure 5.3 shows the Cayley graph obtained from these relator tables.

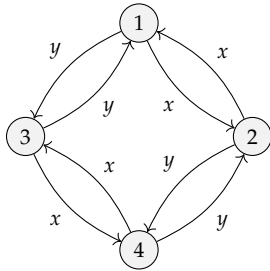


Figure 5.3: The Cayley graph for the group $\langle x, y : x^2, y^2, xyx^{-1}y^{-1} \rangle \cong V_4$

Sometimes when applying this process, we obtain a new deduction that is inconsistent with what we've already got. That is, we might obtain two expressions of the form $ix = j$ and $ix = k$ where $j \neq k$: this is called a **coincidence**. At first sight, this may seem disastrous, and we might be tempted to give up on the whole thing as a bad idea.

But we can salvage the situation by going through all of the tables replacing the larger of j and k with the smaller one. That is, if $j < k$, we rewrite all the tables, replacing any occurrences of k with j , and deleting any rows corresponding to k . If we like, we can renumber everything so that we have a consecutive sequence of integers with no gaps in it. This phenomenon is called **coset collapse**, and after it has happened, we can restart the scanning process from where we left off.

Example 5.68 Let $G = \langle x, y : x^2y, x^3y^2 \rangle$. After a few steps of the Todd–Coxeter algorithm, we get the following situation:

x	x	y	x	x	x	y	y	definition	deduction
1	2	3	1	2	3	4	3	$1x = 2$	
2	3	4	2	3	4			$2x = 3$	$3y = 1$
3	4		3	4				$3x = 4$	$4y = 3$
4		4	4				4		$4y = 2$

At this point we notice that we have a coincidence: $4y = 3$ and $4y = 2$ in the deduction column of the coset table. So before going

any further we rewrite the relator tables, replacing all occurrences of 3 with 2, deleting the third row, and then rewriting all occurrences of 4 as 3. This yields the following situation:

x	x	y	x	x	x	y	y	definition deduction
1	2	2	1	2	2	3	2	$1x = 2$
2	2	3	2	2	3			$2x = 2 \quad 2y = 1$
3		3	3				3	$2x = 3 \quad 3y = 2$

Again, we find a coincidence: $2x = 2$ and $2x = 3$ in the definition column of the coset table. As before, we replace all 3s with 2s and delete the third row of each relator table:

x	x	y	x	x	x	y	y	definition deduction
1	2	2	1	2	2	2	2	$1x = 2$
2	2	2	2	2	2			$2x = 2 \quad 2y = 1$
								$2y = 2$

This yields another coincidence: $2y = 1$ and $2y = 2$, so we replace all 2s with 1s and delete the second row of each relator table:

x	x	y	x	x	x	y	y	definition deduction
1	1	1	1	1	1	1	1	$1x = 1 \quad 1y = 1$

This leaves us with just a single row in both relator tables, from which we conclude that $|G| = 1$. Thus G must be the trivial group.

With a little bit of work, we can amend the basic Todd–Coxeter algorithm to work relative to some finitely-generated subgroup H of the group G we’re interested in. In this case, instead of obtaining the order $|G|$ of the main group, we get the index $|G : H|$ of H in G . And rather than a permutation representation for G , we get a representation of the way G permutes the (right) cosets of H .

The procedure is very similar to that in Examples 5.65, 5.67 and 5.68, but we also have relator tables for the words which generate H . These **subgroup tables** are similar to relator tables in the basic version of the algorithm, but have a single row beginning and ending with 1.

Example 5.69 Let $G = \langle x : x^4 \rangle$ and let $H = \langle x^2 \rangle$. We can see immediately that $G \cong \mathbb{Z}_4$ and that H is an index-2 subgroup which is isomorphic to $\mathbb{Z}_2$. We write down the relator table, the subgroup table and the coset table:

x	x	x	x	x	x	definition deduction
1				1	1	

We make the obvious first move, writing a 2 in all the appropriate

empty cells in the relator table and the subgroup table, and adding the definition $1x = 2$ to the coset table. We add a second row to the relator table, but not to the subgroup table. From the subgroup table we deduce that $2x = 1$, and use this to complete all the tables:

$x \quad x \quad x \quad x$				$x \quad x$		definition deduction	
1	2	1	2	1	2	$1x = 2$	$2x = 1$
2	1	2	1				

The relator table has two rows, which tells us that $|G : H| = 2$. Reading down the first column of the relator table gives us a representation of the generator x as a transposition (12) . With the basic version of the algorithm, this would give us a permutation representative of the generator x of G . This relative version of the Todd–Coxeter algorithm, however, instead tells us how G permutes the (right) cosets of H .

In the next example we apply the Todd–Coxeter algorithm to the dihedral group D_3 relative to the subgroup H generated by one reflection.

Example 5.70 Let $G = \langle x, y : x^3, y^2, (xy)^2 \rangle \cong D_3$ and let $H = \langle y \rangle$. Then after applying the Todd–Coxeter algorithm we obtain the following completed relator tables, subgroup table and coset table:

$x \quad x \quad x$				$y \quad y$		$x \quad y \quad x \quad y$				y
1	2	3	1	1	1	1	2	3	1	1
2	3	1	2	2	3	2	3	2	3	1
3	1	2	3	3	2	3	1	1	2	1

definition deduction	
	$1y = 1$
$1x = 2$	
$2x = 3$	$3x = 1$
	$2y = 3$
	$3y = 2$

This tells us that $|G : H| = 3$, and also that x permutes the right cosets of H by means of a 3-cycle (123) , which maps H to Hx , maps Hx to Hx^2 and maps Hx^2 back to H . Similarly, y permutes the right cosets of H by means of a transposition (23) : this leaves H alone, and transposes Hx and Hx^2 .

Having looked at some illustrative examples, it's time we considered what the various tables and numbers in the algorithm actually mean.

In general, we start with three finite sets: a set X of generators, a set R of relators (such that $G = \langle X : R \rangle$) and a set Y of words which

generate the subgroup $H \leq G$. As the algorithm progresses, we gradually fill up the cells in the relator and subgroup tables with positive integers; these integers correspond to the right cosets of H in G , with 1 representing H itself. This is why the number of integers we end up using gives us the index $|G : H|$ of H in G .

This process is often called **coset enumeration**, for the obvious reason that it lists the cosets of a given subgroup. If the index $|G : H|$ (or, with the simple version of the algorithm, the order $|G|$) is infinite, then the algorithm will not terminate but will just carry on indefinitely. Since we only have a finite amount of time to devote to any given coset enumeration problem, we can't tell whether $|G : H|$ is actually infinite, or just finite but sufficiently large that the algorithm hasn't terminated yet.

In the simple version of the algorithm, the set Y is empty, so the subgroup H is trivial, and the positive integers in the relator tables represent the cosets of this trivial subgroup; that is, the individual elements of G . This, similarly, is why the number of integers we use gives us the order $|G|$ of G .

We can get even more information from the Todd–Coxeter coset enumeration process. To explore this, we need the following definition.

Definition 5.71 Let G be a group and H be a subgroup of G . Then a **right transversal** of H in G is a set of elements of G , one chosen from each right coset of H .

A **left transversal** is defined similarly as a choice of elements from the left cosets of H in G .

To illustrate this, let's look at an example.

Example 5.72 Let $G = D_3$ and $H = \{e, m_1\}$. Then

$$\begin{aligned} H &= Hm_1 = \{e, m_1\}, \\ Hr &= Hm_2 = \{r, m_2\} \\ \text{and } Hr^2 &= Hm_3 = \{r^2, m_3\}, \end{aligned}$$

(see Example 2.18) so $\{e, m_2, r^2\}$ is a right transversal of H in G .

We know from Proposition 2.29 that the cosets of H in G are all the same size, so one way of picturing a transversal is to take all of the cosets, line them up next to each other, and slice through all of them in such a way that we hit a single element in each one.

The Todd–Coxeter algorithm, in addition to counting the number of cosets of H in G , also yields a right transversal for the H in G .

In Example 5.65 we applied the Todd–Coxeter algorithm to the group $\langle x : x^3 \rangle \cong \mathbb{Z}_3$ and found, amongst other things, that $|G| = 3$ as

expected. More generally, what we were doing was making a list of the cosets of the trivial subgroup of G , which coincide with the elements of G itself. There is only one possible transversal in this case: $\{e, x, x^2\}$. We get this transversal recursively from the definition column of the coset table, which in this case consists of $1x = 2$ and $2x = 3$. The first element of the transversal is the identity e , and in general the j th element is given recursively by applying the definitions in the table. So the second element of the transversal is given by the definition $2 = 1x$, which gives $ex = x$, and the third element is given by the definition $3 = 2x$, from which we obtain $xx = x^2$.

In Example 5.67 we constructed a coset table whose definition column consisted of the equations $1x = 2$, $1y = 3$ and $2y = 4$. Following the same procedure, the first element of the transversal is e , the second (given by $2 = 1x$) is x , the third (given by $3 = 1y$) is y , and the fourth (given by $4 = 2y$) is xy . So our transversal is $\{e, x, y, xy\}$ as expected.

More interesting is the relative case where we apply the algorithm relative to a nontrivial subgroup. In Example 5.69 we obtained a single definition $1x = 2$ which we can use to construct the transversal $\{e, x\}$. The first of these is an element of the coset $He = H = \{e, x^2\}$ and the second is an element of the other coset $Hx = \{x, x^3\}$.

The coset table in Example 5.70 contains two definitions, $1x = 2$ and $2x = 3$, which yield the transversal $\{e, x, x^2\}$. Using the isomorphism from the discussion following Proposition 5.28 we can rewrite this as $\{e, r, r^2\}$, and see that it is different to the one given in Example 5.72.

5.B Transversals

THE TRANSVERSALS OBTAINED VIA THE TODD-COXETER ALGORITHM all satisfy an important condition that we're going to rely on during part of our proof of the Nielsen-Schreier Theorem.¹⁹

The recursive way the transversal elements were defined ensures that for any element $x_1x_2 \dots x_3$ in some transversal U , the elements x_1 , x_1x_2 , all the way up to $x_1x_2 \dots x_{n-1}$, are also contained in U .

Definition 5.73 Let G be a group (not necessarily free) generated by some set X , let H be a subgroup of G , and let U be a right transversal of H in G . Then U is a **Schreier transversal**, or has the **Schreier property** if it is **prefix closed**. That is, for any word $x_1x_2 \dots x_n$ in U , the words x_1 , x_1x_2 , all the way up to $x_1x_2 \dots x_{n-1}$, are also in U .

¹⁹ Theorem 5.11, page 143.

In other words, if we delete some substring from the right end of a word in U , what we're left with is also in U .

It's not obvious that any such subgroup has a Schreier transversal, but we'll see in a little while that this actually is the case. First we need to introduce the concept of an ordering on a set:

Definition 5.74 Let $\leq$ be a relation defined on a set S such that:

- P₁ $a \leq a$ (**reflexivity**),
 P₂ if $a \leq b$ and $b \leq a$ then $a = b$ (**antisymmetry**), and
 P₃ if $a \leq b$ and $b \leq c$ then $a \leq c$ (**transitivity**)

for all $a, b, c \in S$. Then $\leq$ is a **partial order** on S .

We say that $\leq$ is a **total order** if it also satisfies the following **trichotomy law** or **comparability** criterion:

- P₄ Either $a \leq b$ or $b \leq a$ for all $a, b \in S$.

For any such partial order $\leq$, there is a corresponding **strict** version $<$, satisfying the following conditions for all $a, b, c \in S$:

- S₁ $a \not\leq a$ (**irreflexivity**),
 S₂ if $a < b$ then $b \not\leq a$ (**asymmetry**), and
 S₃ if $a < b$ and $b < c$ then $a < c$ (**transitivity**)

We say $<$ is a **strict total order** if it also satisfies the trichotomy law:

- S₄ Either $a < b$ or $b < a$ for all $a, b \in S$.

Perhaps the best-known examples of orders are the usual ones $<$ and $\leq$ defined on the real numbers $\mathbb{R}$, and indeed these are the models on which the above definition is based.

For our purposes we need a slight refinement of the above concepts:

Definition 5.75 Let $\leq$ be a total ordering defined on a set S , and let $<$ be the corresponding strict total ordering. If every nonempty subset of S has a least element with respect to $\leq$ and $<$, then we say that S is **well-ordered**, and that both $\leq$ and $<$ are **well-orderings**.

With a bit of thought it should be obvious that any finite set can be well-ordered: arrange the elements of the set in a list and define a strict ordering corresponding to this arrangement. Then the least element is simply the first element on the list. This method also extends to countably infinite sets, but the assertion that *any* set can be well-ordered is equivalent to Zorn's Lemma or the Axiom of Choice. A discussion of the details of axiomatic set theory is beyond the scope of this book, but the statement of both the Axiom and the Lemma are given here for completeness, without further comment.

In particular, we'll use the following strict well-ordering, called variously the **shortlex**, **radix**, **lenlex** or **length-lexicographical** order.

Axiom of Choice Given any collection X of nonempty sets, there exists a function f defined on X (a **choice function**) such that $f(S) \in S$.

Zorn's Lemma Let S be a partially-ordered set such that every totally-ordered, nonempty subset of S has an upper bound. Then S contains at least one maximal element.

Definition 5.76 Let X be a set of generators, equipped with some total order $\leq$, and let $W(X^\pm)$ denote the set of reduced words formed from those generators and their formal inverses. Then the order $\leq$ on X extends to a well-defined order $\leq$ on $W(X^\pm)$, called the **shortlex** or **lenlex** order, defined as follows.

Suppose $u = u_1 \dots u_s$ and $v = v_1 \dots v_t$ be two words in $W(X^\pm)$. Then $u < v$ if and only if:

- (i) either $s < t$
- (ii) or $s = t$ and there exists some $i < s$ such that $u_j = v_j$ for $j \leq i$ but $u_{i+1} < v_{i+1}$.

To see how this works, consider $F_2 \cong \langle x, y : \rangle$. We first need to impose a strict total ordering on the set $X^\pm = \{x, y, x^{-1}, y^{-1}\}$, so let's order the generators x and y alphabetically, and then order their inverses alphabetically too:

$$x < y < x^{-1} < y^{-1}.$$

Now consider the set

$$S = \{x^2, x, x^{-1}, xy, x^{-3}, xyx^{-1}, xyx, y^4\} \subseteq F_2.$$

To put these elements in shortlex order, we first group the words by length, in accordance with condition (i) of Definition 5.76, to get

$$\{x, x^{-1}\} < \{xy, x^2\} < \{x^{-3}, xyx^{-1}, xyx\} < \{y^4\}.$$

Now we use condition (ii) to order words of the same length:

$$x < x^{-1} < x^2 < xy < xyx < xyx^{-1} < x^{-3} < y^4.$$

We can now use the shortlex order to prove that every subgroup of a free group has a Schreier transversal. This is one of the first ingredients in our proof of the Nielsen–Schreier Theorem.

Proposition 5.77 *Let F be a free group with basis X , and let H be a subgroup of F . Then H has a Schreier transversal U in F .*

Proof To form a transversal of H in F , we choose an element from each right coset Hw , where $w \in F$. But to ensure it satisfies the Schreier property, we must be careful about which elements we choose. So, for every coset Hw , select the least element of Hw relative to the shortlex ordering; we claim that this gives a Schreier transversal.

Let $a_1 \dots a_s$ be some element of U . If $a_1 \dots a_{s-1}$ isn't in U then there must exist some word $b_1 \dots b_t a_s$ in the coset $Ha_1 \dots a_{s-1}$ such that $b_1 \dots b_t < a_1 \dots a_{s-1}$. Then $b_1 \dots b_t a_s < a_1 \dots a_{s-1} a_s$ and also $b_1 \dots b_t a_s$ is in the coset $Ha_1 \dots a_{s-1} a_s$.

But this is a contradiction: if $a_1 \dots a_s \in U$ then it must be the least element of the coset $Ha_1 \dots a_s$ relative to the shortlex ordering. Hence $a_1 \dots a_{s-1}$ must be in U after all, and so U is a Schreier transversal. $\square$

In this proof we used the fact (left as an exercise) that for the shortlex ordering $<$ if $v, w \in F$ and $x \in X^\pm$, with $v < w$, then $vx < wx$ and $xv < xw$. In fact, any ordering of F satisfying this property will suffice. Now that we know every subgroup of a free group has a Schreier transversal, let's look at a couple of examples.

Example 5.78 Let $F = F_2 = \langle x, y : \rangle$ and let $H = \langle x^2, y^2, xyx^{-1}y^{-1} \rangle$. In particular, $F/H = \langle x, y : x^2, y^2, xyx^{-1}y^{-1} \rangle \cong V_4$, so $|F : H| = 4$ and thus any transversal of H in F must have four elements. It so happens that

$$U = \{e, x, y, xy\}$$

is a Schreier transversal for H in F . Furthermore, U is the Schreier transversal constructed in Proposition 5.77 since each element of U is the least element in its coset relative to the shortlex ordering.

Sometimes the “obvious” Schreier transversal isn't the optimal one relative to the shortlex ordering, as the following example shows.

Example 5.79 Let $F = F_2 = \langle x, y : \rangle$ and let $H = \langle x^3, y^2, (xy)^2 \rangle$. It so happens that $F/H = \langle x, y : x^3, y^2, (xy)^2 \rangle \cong D_3$, so $|F : H| = 6$ and any transversal of H in F must have six elements.

If $f : X \rightarrow D_3$ defined by $f(x) = r$ and $f(y) = m_1$ then f extends to a unique homomorphism $\tilde{f} : F_2 \rightarrow D_3$ whose kernel is H . We also have

$$\begin{aligned} \tilde{f}(e) &= e, & \tilde{f}(x^2) &= r^2, \\ \tilde{f}(xy) &= m_2, & \text{and } \tilde{f}(x^2y) &= m_3. \end{aligned}$$

We can use these (see Proposition 4.38) to show each of the elements e, x, y, x^2, xy and x^2y lie in different cosets of H in F . Therefore

$$U = \{e, x, y, x^2, xy, x^2y\}$$

is a transversal for H in F . It's also a Schreier transversal, but it's not the minimal Schreier transversal constructed by Proposition 5.77.

To find this shortlex-minimal Schreier transversal, we can list words in F in increasing shortlex order until we've found a representative for each coset, discarding subsequent representatives for a given coset. We then use these elements to form a transversal.

word	e	x	y	x^{-1}	y^{-1}	x^2	xy	xy^{-1}	yx
coset	H	Hx	Hy	Hx^{-1}	Hy	Hx^{-1}	Hxy	Hxy	Hyx

This process yields the transversal

$$V = \{e, x, y, x^{-1}, xy, yx\}$$

which clearly has the Schreier property.

The next step in our journey towards a proof of the Nielsen–Schreier

Theorem is to find generators for the subgroup H , for which we use the shortlex-minimal Schreier transversal provided by Proposition 5.77.

First, though, we introduce another little piece of notation. Suppose that $w \in F$ is some word in a free group F , and that U is a Schreier transversal for a subgroup $H \leq F$. We form the coset Hw and take the intersection $Hw \cap U$. This will consist of a single element which might (but won't necessarily) be equal to w . Denote this element by $\bar{w}$. This process yields a function $\bar{\cdot} : F \rightarrow U$, given by $w \mapsto \bar{w}$ for all $w \in F$.

This **bar function** has a number of properties which we state here without further proof. It's **idempotent**, in the sense that $\bar{\bar{w}} = \bar{w}$ for all $w \in F$. Also, $Hw = H\bar{w}$ for all $w \in F$, and $\bar{w} = w$ if and only if $w \in U$.

Let F be a free group generated by some basis X , let H be a subgroup of F , and let U be a Schreier transversal for H in F . In the next part of this discussion we'll need the set

$$Y = \{ux\bar{u}\bar{x}^{-1} : u \in U, x \in X^{\pm}\}.$$

It's worth thinking a little bit about what this set Y actually means. Suppose that u is an element of the Schreier transversal U , and that x is either a generator, or the inverse of a generator, in X . Then ux is clearly a word in F , and $\bar{u}\bar{x}$ is the corresponding transversal element in U , and may or may not be equal to ux . If $ux = \bar{u}\bar{x}$, which will happen exactly when $ux \in U$, then the element $ux\bar{u}\bar{x}^{-1}$ will be equal to the identity $e \in F$. So $ux\bar{u}\bar{x}^{-1}$ tells us whether ux is also in U .

Now consider the element $\overline{\bar{u}\bar{x}x^{-1}}$. This is in the same coset as $\bar{u}\bar{x}x^{-1}$, which is in the same coset as $uxx^{-1} = u$. Hence

$$\overline{\bar{u}\bar{x}x^{-1}} = u. \quad (5.2)$$

We'll use this fact a couple of times later, but first let's construct the set Y in a specific case to see how it works.

Example 5.80 Let $F = F_2 = \langle x, y : \rangle$, let $H = \langle x^3, y^2, (xy)^2 \rangle$, and let

$$U = \{e, x, y, x^{-1}, xy, yx\}$$

be the shortlex-minimal Schreier transversal obtained via Proposition 5.77. We construct the set $Y = \{ux\bar{u}\bar{x}^{-1} : u \in U, x \in X^{\pm}\}$ by means of the following table:

	x	y	x^{-1}	y^{-1}
e	e	e	e	y^{-2}
x	x^3	e	e	$xy^{-2}x^{-1}$
y	e	y^2	$yx^{-1}y^{-1}x^{-1}$	e
x^{-1}	e	$x^{-1}yx^{-1}y^{-1}$	x^{-3}	$x^{-1}y^{-1}x^{-1}y^{-1}$
xy	xyx^{-1}	xy^2x^{-1}	$xyx^{-2}y^{-1}$	e
yx	$yx^2y^{-1}x^{-1}$	$yx yx$	e	$yx y^{-1}x^{-1}$

Here the column headings are elements x of $X^\pm$ while the row headings are elements u of the transversal U , and the entries in the table are the corresponding words $ux\bar{u}x^{-1}$.

For example, we calculate $(xy)x^{-1}\overline{(xy)x^{-1}}^{-1}$ by first concatenating xy and x^{-1} to get xyx^{-1} . Then we find the appropriate transversal element $\overline{xyx^{-1}} = yx$, invert it to get $x^{-1}y^{-1}$, append it to xyx^{-1} , and perform any necessary cancellations to get $xyx^{-2}y^{-1}$.

The set Y , then, comprises all of these elements:

$$Y = \{e, y^2, y^{-2}, x^3, xyx^{-1}, x^{-3}, xy^2x^{-1}, xy^{-2}x^{-1}, yxyx, \\ yxy^{-1}x^{-1}, yx^{-1}y^{-1}x^{-1}, x^{-1}yx^{-1}y^{-1}, x^{-1}y^{-1}x^{-1}y^{-1}, \\ xyx^{-2}y^{-1}, yx^2y^{-1}x^{-1}\}$$

We won't require that U is minimal with respect to the shortlex ordering in the rest of this discussion, but the shortlex ordering is useful because it enables us to construct a Schreier transversal.

The idea is that this process yields a set of generators for the subgroup H , but looking at the table constructed in the above example, we see that there are a number of redundant or trivial elements on the list. In particular, there are several instances of the identity e : this occurs when $ux \in U$ so that $ux = \bar{u}x$. Also, there are a number of occasions when both an element and its inverse occur. In fact, the nontrivial elements in the right half of the table are exactly the inverses of those in the left half of the table. So, in general, we can halve the amount of work we have to do by only constructing the left half of the table.

Now define

$$Z = \{ux\bar{u}x^{-1} : u \in U, x \in X, ux \notin U\}.$$

This will yield only the nontrivial elements we want. To prove that these are the only redundancies we need the following lemma.

Lemma 5.81 *Let*

$$Z^{-1} = \{z^{-1} : z \in Z\}$$

$$\text{and } \bar{Z} = \{ux\bar{u}x^{-1} : u \in U, x \in X^{-1}, ux \notin U\}.$$

Then $Z^{-1} = \bar{Z}$ and $Y = Z \cup Z^{-1} \cup \{e\}$.

Proof Let $w \in Z^{-1}$, so $w = (ux\bar{u}x^{-1})^{-1} = \bar{u}xx^{-1}u^{-1}$ for some $u \in U$ and $x \in X$. Then by (5.2) we have $u = \overline{\bar{u}xx^{-1}}$ so $w = \bar{u}xx^{-1}\overline{\bar{u}xx^{-1}}^{-1}$, which is an element of $\bar{Z}$, since we can rewrite it as $w = vx^{-1}\overline{vx^{-1}}^{-1}$ where $v = \bar{u}x$. Hence $Z^{-1} \subseteq \bar{Z}$.

Now let $w = ux^{-1}\bar{u}x^{-1} \in \bar{Z}$, so $w^{-1} = \overline{ux^{-1}}xu^{-1}$. Writing $v = \overline{ux^{-1}}$ we see that $\bar{v}x = u$ by (5.2) and hence $w^{-1} = vx\bar{v}x^{-1} \in Z$, so $w \in Z^{-1}$. Therefore $\bar{Z} \subseteq Z^{-1}$ and so $\bar{Z} = Z^{-1}$ as required.

The union $Z \cup Z^{-1}$ contains all the nontrivial elements of Y , so including e as well gives us the entirety of Y . $\square$

The significance of Z is that its elements generate H :

Lemma 5.82 *Let F be a free group generated by some basis X , let H be a subgroup of F , and let U be a Schreier transversal for H in F . The set $Z = \{ux\bar{u}\bar{x}^{-1} : u \in U, x \in X, ux \notin U\}$ generates H .*

Proof Let $h \in H$, and suppose that $h = x_1 \dots x_n$ where $x_1, \dots, x_n \in X^\pm$. We now define a sequence of transversal elements $u_1, \dots, u_{n+1}$ by setting $u_1 = e$ and $u_{i+1} = \overline{x_1 \dots x_i}$ for $1 \leq i \leq n$. Since $h \in H$ we have $u_{n+1} = \overline{x_1 \dots x_n} = \bar{h} = e$. Also, the elements u_i satisfy a recursive relation $u_{i+1} = \overline{u_i x_i}$.

Define a sequence of elements $y_1, \dots, y_n \in Y$ by setting

$$y_i = u_i x_i \overline{u_i x_i}^{-1} = u_i x_i u_{i+1}^{-1}.$$

Then

$$\begin{aligned} y_1 \dots y_n &= (u_1 x_1 u_2^{-1})(u_2 x_2 u_3^{-1}) \dots (u_n x_n u_{n+1}^{-1}) \\ &= u_1 x_1 \dots x_n u_{n+1}^{-1} \\ &= u_1 h u_{n+1}^{-1} \\ &= e h e \\ &= h. \end{aligned}$$

This tells us that any element $h \in H$ can be written as a product of elements of Y , each of which is in turn a word in elements of Z and their inverses. Therefore Z generates H as claimed. $\square$

At this point, we've found a set Z of generators for H (which we will sometimes refer to as **Schreier generators**), but we still need to show that H is freely generated by those elements. The way we do this is by means of the following lemma.

Lemma 5.83 *Let $u \in U$ and $x \in X^\pm$ such that $ux\bar{u}\bar{x}^{-1} \in Z^\pm$. Then x doesn't cancel in the product $ux\bar{u}\bar{x}^{-1}$. Furthermore, if $v \in U$ and $y \in X^\pm$ such that $vy\bar{v}\bar{y}^{-1} \in Z^\pm$ and $vy\bar{v}\bar{y}^{-1} \neq (ux\bar{u}\bar{x}^{-1})^{-1}$ then in the product $ux\bar{u}\bar{x}^{-1}vy\bar{v}\bar{y}^{-1}$ neither the first x nor the first y cancels.*

Proof If x cancels in $ux\bar{u}\bar{x}^{-1}$, then it either does so in ux or in $x\bar{u}\bar{x}^{-1}$. In the former case, let $u = a_1 \dots a_m$ with $a_m^{-1} = x$. Then $ux = a_1 \dots a_{m-1}$, which is in U by the Schreier property, so $ux = \bar{u}\bar{x}$ and hence $ux\bar{u}\bar{x}^{-1} = e$, contradicting our hypothesis that $ux\bar{u}\bar{x}^{-1} \in Z^\pm$.

The other possibility is that x cancels in $x\bar{u}\bar{x}^{-1}$. For this to happen, we need $\bar{u}\bar{x} = b_1 \dots b_n$ with $x = b_n$ so that $\bar{u}\bar{x}x^{-1} = b_1 \dots b_{n-1} \in U$ and hence $\bar{u}\bar{x} = ux$, which also contradicts the assertion that $ux\bar{u}\bar{x}^{-1} \in Z^\pm$.

Now consider the product $ux\bar{u}\bar{x}^{-1}vy\bar{v}\bar{y}^{-1}$. By the above discussion, x can't cancel with either u or $\bar{u}\bar{x}^{-1}$, and y can't cancel with either v or $\bar{v}\bar{y}^{-1}$, so if x cancels at all, then either all of $\bar{u}\bar{x}^{-1}$ cancels with (some or all of) v , or all of v cancels with (some or all of) $\bar{u}\bar{x}^{-1}$.

Suppose that all of $\bar{u}\bar{x}^{-1}$ cancels with some of v . Then $\bar{u}\bar{x} = a_1 \dots a_m$ and $v = b_1 \dots b_n$ with $m \leq n$ and $a_i = b_i$ for $1 \leq i \leq m$.

If $m < n$ then x cancels with b_{m+1} in the sense that $x = b_{m+1}^{-1}$, and thus $\bar{u}\bar{x}x^{-1} = b_1 \dots b_{m+1} \in U$ (by the Schreier property). Hence $\bar{u}\bar{x}x^{-1} = u$ by (5.2), so $\bar{u}\bar{x} = ux$, contradicting the hypothesis that $ux \notin U$.

If, on the other hand, $m = n$ then $\bar{u}\bar{x} = v$ and $x = y^{-1}$, so $vy = \bar{u}\bar{x}x^{-1}$, which means that $\bar{v}\bar{y} = \bar{u}\bar{x}x^{-1} = u$. Hence $ux\bar{u}\bar{x}^{-1} = \bar{v}\bar{y}\bar{y}^{-1}v^{-1} = (v\bar{y}\bar{v}\bar{y}^{-1})^{-1}$, which contradicts our hypothesis.

A similar argument shows that all of v can't cancel with (some or all of) $\bar{u}\bar{x}^{-1}$, at which point the lemma is proved. $\square$

Corollary 5.84 *Let $y_i = u_i x_i \bar{u}_i \bar{x}_i^{-1}$ in $Z^\pm$ for $1 \leq i \leq m$ and suppose that $y_i \neq y_{i+1}^{-1}$ for $1 \leq i < m$. Then in the product $y_1 \dots y_m$ none of the x_i terms cancel, so $y_1 \dots y_m \neq e$.*

At long last, we are now ready to prove the Nielsen–Schreier Theorem.

Proof of Theorem 5.11 By Proposition 5.77 we can find a Schreier transversal U for H in F such that $|U| = |F : H| = s$.

By Lemmas 5.81 and 5.82 we can use this transversal to construct a generating set Z for H .

To show that H is free on Z , we just have to show that distinct reduced words in $Z^\pm$ represent distinct elements of H . Suppose that $b_1 \dots b_m = c_1 \dots c_n$ in F with $b_i, c_i \in Z^\pm$, and $b_i \neq b_{i+1}^{-1}$ for $1 \leq i < m$ and $c_1 \neq c_{i+1}$ for $1 \leq i < n$. Then $b_1 \dots b_m c_n^{-1} \dots c_1^{-1} = 1$, so by Corollary 5.84 we must have $b_m = c_n$. We can show by induction that $b_i = c_i$ for $1 \leq i \leq m$ and that $m = n$, so H is free on Z .

To calculate the rank of H , we need to work out $|Z|$. In other words, we need to know how many $u \in U$ and $x \in X$ satisfy $ux = \bar{u}\bar{x}$.

Let $a_1 \dots a_m \in U \setminus \{e\}$. If $a_m = x \in X$ then we have $ux = \bar{u}\bar{x}$ with $u = a_1 \dots a_{m-1}$ since U is a Schreier transversal. On the other hand, if $a_m = x^{-1} \in X^{-1}$ then $ux = \bar{u}\bar{x}$ with $u = a_1 \dots a_m$. So in either case there is a trivial Schreier transversal generator associated with u .

Conversely, if $ux\bar{u}\bar{x}^{-1}$ is trivial then x cancels either with the last letter of u or the first letter of $\bar{u}\bar{x}^{-1}$ so it must occur in one of the two ways just described. Hence the number of trivial elements in Z is $|U \setminus \{e\}| = n - 1$.

So $|Z| = nr - (n - 1) = n(r - 1) + 1$ as claimed. $\square$

Theorem 5.11 (The Nielsen–Schreier Theorem) *Let F be a free group, and let H be a subgroup of F . Then H is itself a free group. Furthermore, if the index $|F:H| = s$ of H in F and $\text{rank}(F) = r$ are both finite, then*

$$\text{rank}(H) = (r - 1)s + 1.$$

During this section we've been primarily concerned with subgroups of free groups, but it turns out that we can use the information provided by the Todd–Coxeter algorithm to write down a presentation of a finite-index subgroup H of any finitely-presented group G .

Suppose $G = \langle X : R \rangle$ and $H = \langle Y \rangle \leq G$. Then the Todd–Coxeter algorithm provides us with a Schreier transversal U for H in G . We then use the process described earlier (in Lemmas 5.81 and 5.82) to construct a generating set Z for H : they are simply the nontrivial words $ux\bar{u}x^{-1}$ where $x \in X$ and $u \in U$.

Next, we must form a set of relators for H expressed in terms of the generators X , and finally we have to rewrite those relators in terms of the generators in Z . The following proposition gives us the key to the first of these tasks.

Proposition 5.85 *Let $G = \langle X : R \rangle$, let H be a subgroup of G , and let U be a Schreier transversal for H in G . Then $H = \langle Z : S \rangle$ where*

$$Z = \{ux\bar{u}x^{-1} : u \in U, x \in X, ux \notin U\}$$

and S is the set of elements of

$$T = \{uru^{-1} : u \in U, r \in R\}$$

rewritten as words in $Z^\pm$.

Proof We prove this by means of a slightly more general result.

Let F be a group, let K be a subgroup of F , and let V be a Schreier transversal of K in F . Let Q be a subset of F such that $\langle Q^F \rangle \subseteq K$, where $\langle Q^F \rangle$ denotes the normal closure of Q in F .

We claim that $\langle Q^F \rangle$ is also the normal closure in K of the set

$$P = \{vqv^{-1} : v \in V, q \in Q\}.$$

To see this, recall that $\langle Q^F \rangle$ is generated by elements wqw^{-1} where $w \in F$ and $q \in Q$. Any $w \in F$, however, can be written in the form kv for some $k \in K$ and $v \in V$. So $\langle Q^F \rangle$ is generated by all elements $kvqv^{-1}k^{-1} = k(vqv^{-1})k^{-1}$, which are all conjugates of elements of P with elements of K , so $\langle Q^F \rangle = \langle P^K \rangle$ as claimed.

We can now apply this to the specific problem of finding relators for H in terms of the generators in Z .

Set $F = F(X)$ and let K be the preimage $f^{-1}(H)$ in F , where $f: F \rightarrow G = F/\langle R^F \rangle$ is the quotient homomorphism obtained by factoring F by the relators R .

(By Proposition 4.37, $f^{-1}(H)$ must be a subgroup of F , and in this case we also know, by the Nielsen–Schreier Theorem, that it's free.)

We have a Schreier transversal U of H in G , but we can also regard this as a Schreier transversal for K in F . To see this, observe that U is

formed by taking a carefully-chosen element from each coset of H in G . Passing to the free group F , these same transversal elements can also be seen as belonging to the cosets of K in F .

Therefore K is freely generated by the set Z . The relator set R is obviously a subset of F , and its normal closure $\langle R^F \rangle$ lies in K , so by the more general discussion above we know that $\langle R^F \rangle = \langle T^K \rangle$.

Rewriting the elements of T in terms of the generators in Z (which we can often just do by inspection) we get the set S .

All that remains is to note that $\langle Z : S \rangle \cong K / \langle R^F \rangle \cong H$. □

So, in practice we write down the set

$$T = \{uru^{-1} : u \in U, r \in R\}$$

and then rewrite these in terms of the generators in Z , either by inspection or by some other method such as the recursive procedure outlined in the proof of Lemma 5.82. This is all best illustrated by a couple of examples.

Example 5.86 In Example 5.69 we applied the Todd–Coxeter algorithm to the group $G = \langle x : x^4 \rangle \cong \mathbb{Z}_4$ and the subgroup $H = \langle x^2 \rangle$. The resulting Schreier transversal is $U = \{e, x\}$, and we can use this to construct the Schreier generators $ux\bar{u}x^{-1}$ for H in G . The only nontrivial such generator is x^2 so we have $Z = \{x^2\}$. Let $z = x^2$ be the only element of this set.

Next we calculate the elements uru^{-1} of the set T , which in this case gives $T = \{x^4\}$. We can immediately rewrite this in terms of the Schreier generator $z = x^2$, giving $x^4 = z^2$ and the presentation

$$H = \langle z : z^2 \rangle \cong \mathbb{Z}_2$$

as expected.

It's illuminating to use the recursive method from Lemma 5.82. We take the relator $x^4 = x_1x_2x_3x_4$, setting $x_1 = x_2 = x_3 = x_4 = x$. Next we recursively calculate the transversal elements

$$\begin{aligned} u_1 &= e, & u_2 &= \bar{x}_1 = \bar{x} = x, \\ u_3 &= \bar{x}_1\bar{x}_2 = \bar{x}^2 = e, & u_4 &= \bar{x}_1\bar{x}_2\bar{x}_3 = \bar{x}^3 = x, \\ u_5 &= \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4 = \bar{x}^4 = e. \end{aligned}$$

We use these to calculate the elements $y_1, \dots, y_4$:

$$\begin{aligned} y_1 &= u_1x_1u_2^{-1} = exx^{-1} = e, & y_2 &= u_2x_2u_3^{-1} = xxe = x^2 = z, \\ y_3 &= u_3x_3u_4^{-1} = exx^{-1} = e, & y_4 &= u_4x_4u_5^{-1} = xxe = x^2 = z. \end{aligned}$$

So $x^4 = ezez = z^2$ as expected, which gives us the presentation $H = \langle z : z^2 \rangle$ again.

Next we study the subgroup of D_3 generated by a single reflection.

Example 5.87 Let $G = \langle X : R \rangle = \langle x, y : x^2, y^2, (xy)^3 \rangle \cong D_3$ and $H = \langle x \rangle \cong \mathbb{Z}_2$. Applying the Todd–Coxeter algorithm tells us that $|G : H| = 3$ and gives us a Schreier transversal $U = \{e, y, yx\}$, which we can use to write down the Schreier generators

$$z_1 = x, \quad z_2 = y^2, \quad z_3 = yxyx^{-1}y^{-1}, \quad z_4 = yx^2y^{-1}.$$

The set $T = \{uru^{-1} : u \in U, r \in R\}$ consists of the words

$$x^2, y^2, (xy)^3, yx^2y^{-1}, (yx)^3, yx^2yxyxyx^{-1}y^{-1}.$$

Some of these can be immediately written in terms of the Schreier generators $z_1, \dots, z_4$:

$$x^2 = z_1^2, \quad y^2 = z_2, \quad yx^2y^{-1} = z_4.$$

For the rest, we use the procedure outlined in Lemma 5.82. Starting with $(xy)^3 = xyxyxy$ and setting $x_1 = x_3 = x_5 = x$ and $x_2 = x_4 = x_6 = y$, we obtain the transversal elements

$$\begin{aligned} u_1 &= e, & u_2 &= \bar{x}_1 = \bar{x} = e, \\ u_3 &= \bar{x}_1\bar{x}_2 = \bar{xy} = y, & u_4 &= \bar{x}_1\bar{x}_2\bar{x}_3 = \bar{xyx} = yx, \\ u_5 &= \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4 = \bar{xyxy} = y, & u_6 &= \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4\bar{x}_5 = \bar{xyxyx} = yx, \\ u_7 &= \bar{x}_1\bar{x}_2\bar{x}_3\bar{x}_4\bar{x}_5\bar{x}_6 = e. \end{aligned}$$

Using these, we can write down the elements $y_1, \dots, y_6$ as follows:

$$\begin{aligned} y_1 &= u_1x_1u_2^{-1} = exe = x = z_1, & y_2 &= u_2x_2u_3^{-1} = eyy^{-1} = e, \\ y_3 &= u_3x_3u_4^{-1} = yxx^{-1}y^{-1} = e, & y_4 &= u_4x_4u_5^{-1} = yxyx^{-1}y^{-1} = z_3, \\ y_5 &= u_5x_5u_6^{-1} = yx^2y^{-1} = z_4, & y_6 &= u_6x_6u_7^{-1} = yx^2y^{-1} = z_2. \end{aligned}$$

Hence $(xy)^3 = (x)(yxyx^{-1}y^{-1})(yx^2y^{-1})(y^2) = z_1z_3z_4z_2$. Following the same process, we get

$$\begin{aligned} (yx)^3 &= (yxyx^{-1}y^{-1})(yx^2y^{-1})(y^2)(x) = z_3z_4z_2z_1, \\ yx^2yxyxyx^{-1}y^{-1} &= (yx^2y^{-1})(y^2)(x)(yxyx^{-1}y^{-1}) = z_4z_2z_1z_3. \end{aligned}$$

This gives us the presentation

$$H = \langle z_1, z_2, z_3, z_4 : z_1^2, z_2, z_4, z_1z_3z_4z_2, z_3z_4z_2z_1, z_4z_2z_1z_3 \rangle.$$

Immediately we can see that two of the generators, namely z_2 and z_4 correspond to trivial relators and can therefore be discarded, yielding

$$H = \langle z_1, z_3 : z_1^2, z_1z_3, z_3z_1, z_1z_3 \rangle.$$

This can be further simplified, either with carefully-chosen Tietze transformations or just by inspection, to give

$$H = \langle z : z^2 \rangle \cong \mathbb{Z}_2.$$

5.C Triangles, braids and reflections

IN THIS SECTION we will look briefly at a few classes of groups that can be defined by presentations of particular forms, but which have interesting geometric interpretations.

We start with the dihedral group D_n , which can be expressed by the presentation

$$\langle x, y : x^2 = y^2 = (xy)^n = e \rangle.$$

This group has an interpretation as the symmetry group of the regular n -gon. Both generators x and y represent reflections, all of which have order 2, and the product xy represents a $\frac{2\pi}{n}$ rotation, which has order n . If we choose x and y to be the right reflections, then we can generate the entirety of D_n just using them. In general there are many valid choices, but setting $x = m_1$ and $y = m_2$ works.

The symmetric group S_n has a presentation

$$\left\langle x_1, \dots, x_{n-1} : \begin{array}{ll} x_i^2 = e & 1 \leq i \leq n-1 \\ (x_i x_{i+1})^3 = e & 1 \leq i \leq n-2 \\ x_i x_j = x_j x_i & 1 \leq i < j \leq n-1, \quad j-i > 1 \end{array} \right\rangle.$$

(The presentation for $D_3 \cong S_3$ is a special case of this.) Again, we have a group whose generators all have order 2, and where products of pairs of generators all have finite order.

A carefully chosen, compact notation can be tremendously useful in seeing to the heart of a problem, and to that end we now introduce the following scheme for describing finitely generated groups whose generators all have order 2.

Definition 5.88 Given a finitely-generated group

$$G = \langle x_1, \dots, x_n : x_i^2 = e, (x_i x_j)^{k_{ij}} = e \text{ for } 1 \leq i < j \leq n \rangle,$$

we construct a graph as follows: Represent each generator x_i by a vertex, and connect two vertices by an edge labelled k if the product of the corresponding generators has order k . To simplify the diagram, we omit the label if $k = 2$ or 3, and omit the edge as well if $k = 2$.

The resulting graph is called a **Coxeter diagram** or **Coxeter graph**;²⁰ a group determined by such a presentation is called a **Coxeter group**.

Given two generators x and y of order 2, if their product xy also has order 2, this is equivalent to saying that x and y commute. We'll view this as the default situation, which is why we omit the edges between vertices representing commuting generators. Also, it turns out that many of the interesting cases mostly have products of order 3, so we omit those labels to simplify the notation.

A bas Euclide! Mort aux triangles!
(Down with Euclid! Death to triangles!)

— Jean Dieudonné (1906–1992)

Although it has been proved that every braid can be deformed into a similar normal form the writer is convinced that any attempt to carry this out on a living person would only lead to violent protests and discrimination against mathematics. He would therefore discourage such an experiment.

— Emil Artin (1898–1962),
Theory of braids, *Annals of Mathematics* 48 (1947) 101–126

²⁰ They are named after the British mathematician HSM Coxeter (1907–2003). A very similar notation was independently devised by the Russian mathematician Eugenii Dynkin (1924–2014), and is particularly useful in the study and classification of Lie groups and Lie algebras. For this reason, these diagrams are also sometimes called **Dynkin diagrams** or **Coxeter–Dynkin diagrams**.

Example 5.89 The dihedral group D_m is represented by the Coxeter graph

or

for D_3 , and

for $D_2 = V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

Example 5.90 The symmetric group S_n is represented by the Coxeter diagram

with n vertices.

²¹ JE Humphreys, *Reflection Groups and Coxeter Groups*, Cambridge Studies in Advanced Mathematics 29, Cambridge University Press (1990).

²² As noted before, in the discussion of finite simple groups of Lie type, we have a slight notational awkwardness to contend with here. We will again use a sans serif typeface to denote these groups and distinguish them from other similarly-named ones. In particular, we distinguish between the alternating group A_n and the Coxeter group A_n (which is isomorphic to the symmetric group S_{n+1}), and between the dihedral group D_n (which in this scheme is denoted $I_2(n)$) and the Coxeter group D_n .

A natural question to ask is which presentations (and which Coxeter diagrams) correspond to finite groups. The classification theorem is reasonably straightforward, but the details would involve us in too much of a digression here, and so the interested reader is directed to the book by James Humphreys.²¹ The complete list of finite Coxeter groups is given in Table 5.3. The list comprises four infinite families (A_n , B_n , D_n and $I_2(m)$), and seven one-off extra cases: the **exceptional** Coxeter groups (E_6 , E_7 , E_8 , F_4 , G_2 , H_3 and H_4).²² In this notation, the subscript denotes the **rank** (the number of generators) of the group.

All of these groups arise as symmetry groups of geometric objects in Euclidean spaces, specifically polygons, polyhedra or their higher-dimensional analogues **polytopes** (or **polychora**). The group $A_3 \cong S_4$ describes the symmetries of the regular tetrahedron; more generally A_n is the symmetry group of the n -**simplex**.

The group B_3 is the symmetry group of the cube and the octahedron; B_4 of their analogues in 4 dimensions (the **hypercube** or **tesseract** and the **hyperoctahedron**); and in general B_n is the symmetry group of the n -cube and n -octahedron (or n -dimensional **cross polytope**).

The groups D_n represent the symmetries of a family of semiregular polytopes called n -**demicubes**. These are obtained by deleting every other vertex from the n -cube; in particular the 3-demicube is a tetrahedron and the 4-demicube is the hyperoctahedron.

The exceptional groups E_6 , E_7 and E_8 describe the symmetries of three semiregular **Gosset polytopes** first discovered by the British barrister and amateur mathematician Thorold Gosset (1869–1962).

The exceptional group F_4 is the symmetry group of a regular 4-dimensional polytope called the 24-**cell**, which consists of 24 regular

Type		Coxeter diagram	Order of group	Polytopes
A_n	$n \geq 1$	$\bullet \text{---} \bullet \text{---} \cdots \text{---} \bullet$	$(n+1)!$	n -simplex
B_n	$n \geq 2$	$\bullet \text{---} \bullet \text{---} \cdots \text{---} \bullet \overset{4}{\text{---}} \bullet$	$2^n n!$	n -cube, n -cross
D_n	$n \geq 4$	$\bullet \text{---} \bullet \text{---} \cdots \text{---} \bullet \begin{array}{l} \nearrow \bullet \\ \searrow \bullet \end{array}$	$2^{n-1} n!$	n -demicube
E_6		$\begin{array}{c} \bullet \\ \\ \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \end{array}$	51 840	Gosset polytope 2_{21}
E_7		$\begin{array}{c} \bullet \\ \\ \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \end{array}$	2 903 040	Gosset polytope 3_{21}
E_8		$\begin{array}{c} \bullet \\ \\ \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \text{---} \bullet \end{array}$	696 729 600	Gosset polytope 4_{21}
F_4		$\bullet \text{---} \bullet \overset{4}{\text{---}} \bullet \text{---} \bullet$	1 152	24-cell
G_2		$\bullet \overset{6}{\text{---}} \bullet$	12	regular hexagon
H_3		$\bullet \overset{5}{\text{---}} \bullet \text{---} \bullet$	120	dodecahedron, icosahedron
H_4		$\bullet \overset{5}{\text{---}} \bullet \text{---} \bullet \text{---} \bullet$	14 400	120-cell, 600-cell
$I_2(m)$	$m \geq 2$	$\bullet \overset{m}{\text{---}} \bullet$	$2m$	regular m -gon

Table 5.3: Finite Coxeter groups

octahedra glued together along their faces.

The group H_3 describes the symmetries of the dodecahedron and the icosahedron, while H_4 is the symmetry group of two regular 4-dimensional polytopes: the 120-cell or **hyperdodecahedron** (which consists of 120 regular dodecahedra glued together along their faces) and the 600-cell or **hypericosahedron** (which has 600 regular tetrahedral hyperfaces).

The groups $I_2(m)$ are isomorphic to the dihedral groups D_m . For complicated reasons that need not concern us here (but which relate to Lie groups and Lie algebras) the case $m = 6$ is sometimes treated as a special case, and denoted G_2 ; also $I_2(3) \cong A_2$ and $I_2(4) \cong B_2$.

Summary

In this chapter we developed the notion of a **presentation** of a group, as well as some powerful technical machinery to study groups defined in this way. We began by examining the construction of the infinite cyclic group $\mathbb{Z}_n$ from all possible powers of a given generator t , and then making the finite cyclic group $\mathbb{Z}_n$ by imposing a constraint $t^n = 1$ on this generator. In order to generalise this approach to larger generating sets, we studied formal reduced strings or **words**²³ in these generating symbols together with their formal inverses. By concatenating words of this type (and performing any necessary cancellations of generators with adjacent formal inverses) we obtained an associative binary operation on the set of such finite reduced words. The empty word e acts as an identity element, and each reduced word has a formal inverse obtained by generalising the usual result $(g * h)^{-1} = h^{-1} * g^{-1}$. This enabled us to construct first the **free group** F_2 with two generators,²⁴ and then to generalise this to the free group on an arbitrary (possibly infinite) set of generators; the number of generators is called the **rank** of the group.²⁵ We then discussed a more abstract definition²⁶ in terms of an adjointness condition, and proved that these two definitions are equivalent.²⁷ Free groups are determined by the cardinality of their generating sets: if we have two sets X_1 and X_2 such that a bijection exists between them, then the free groups $F(X_1)$ and $F(X_2)$ are isomorphic.²⁸ In particular, $F_m \cong F_n$ if and only if $m = n$.²⁹

Next we studied the order of elements in a free group. First we introduced the notion of a **cyclically reduced** word³⁰ and used this to show that free groups are **torsion free**; that is, they have no nontrivial elements of finite order.³¹ Free groups are about as nonabelian as they can be: two elements commute if and only if they are both powers of the same element.³²

Having defined free groups and studied some of their properties, we then introduced the notion of a **presentation**,³³ which we defined in terms of the quotient of a free group by the **normal closure**^{34,35} of a set of carefully-chosen words, which we called **relators** (equivalently we can replace relators with defining equations called **relations**). If a group can be represented in this way using a finite set of generators and relators then we say it's **finitely presented**. Group presentations satisfy a similar universal mapping property³⁶ to the one satisfied by free groups.³⁷ Usefully, it turns out that every group can be described by means of a suitable presentation,³⁸ although not necessarily by a finite presentation.

²³ Definition 5.3, page 137.

²⁴ Proposition 5.2, page 136.

²⁵ Definition 5.4, page 137.

²⁶ Definition 5.5, page 138.

²⁷ Proposition 5.6, page 139.

²⁸ Proposition 5.7, page 140.

²⁹ Corollary 5.8, page 141.

³⁰ Definition 5.9, page 142.

³¹ Proposition 5.10, page 142.

³² Proposition 5.12, page 143.

³³ Definition 5.16, page 147.

³⁴ Definition 5.14, page 146.

³⁵ Proposition 5.15, page 146.

³⁶ Proposition 5.17, page 147.

³⁷ Definition 5.5, page 138.

³⁸ Proposition 5.18, page 148.

Given presentations for two groups $G = \langle X : R \rangle$ and $H = \langle Y : S \rangle$, we can easily write down a presentation $\langle X \cup Y : R \cup S \cup [X, Y] \rangle$ for their direct product $G \times H$.³⁹ Here, $[X, Y]$ denotes the set of commutators $xyx^{-1}y^{-1}$ of generators $x \in X$ and $y \in Y$. This led us to the notions of a **free product**⁴⁰ $G * H$ and an **amalgamated free product**⁴¹ $G *_K H$. In particular, the free product $F(X) * F(Y)$ is equal to the free group $F(X \cup Y)$.⁴²

³⁹ Proposition 5.22, page 149.

⁴⁰ Definition 5.23, page 151.

⁴¹ Definition 5.26, page 151.

⁴² Proposition 5.24, page 151.

We then looked at some presentations for groups we met in earlier chapters. The dihedral groups D_n ,⁴³ the quaternion group Q_8 ,⁴⁴ the additive group $\mathbb{Q}$ of rational numbers,⁴⁵ and the symmetric groups S_n .⁴⁶ In particular, the presentation obtained for the dihedral groups allowed us to prove a useful classification result for groups of order $2p$ with p prime: any such group must be isomorphic to either the cyclic group $\mathbb{Z}_{2p}$ or the dihedral group D_p .⁴⁷ We also introduced the **dicyclic groups** Dic_n which in some sense generalise the quaternion group Q_8 .⁴⁸

⁴³ Proposition 5.28, page 152.

⁴⁴ Proposition 5.30, page 155.

⁴⁵ Proposition 5.32, page 156.

⁴⁶ Proposition 5.33, page 157.

⁴⁷ Proposition 5.29, page 153.

⁴⁸ Example 5.31, page 155.

A discussion of the different presentations obtained for the isomorphic groups D_3 and S_3 led us to formulate certain simple operations, called **Tietze transformations**.⁴⁹ If two presentations are related by a finite sequence of these transformations, they determine isomorphic groups;⁵⁰ furthermore any presentation for a given group can be turned into any other presentation for the same group by a finite sequence of these operations.⁵¹ In practice it is not always easy to find such a sequence: it reduces to the **word problem**, which is in general unsolvable.

⁴⁹ Definition 5.36, page 160.

⁵⁰ Proposition 5.35, page 159.

⁵¹ Proposition 5.41, page 163.

We then investigated the abelian case, obtaining concrete⁵² and abstract⁵³ definitions for the notion of a **free abelian group**, which groups (at least in the finite rank case) we then found to be isomorphic to direct products $\mathbb{Z}^n$ of copies of the group of integers.⁵⁴ We then proved a result of Richard Dedekind (1831–1916) that says that any subgroup of a free abelian group of rank r must also be a free abelian group of rank at most r .⁵⁵

⁵² Definition 5.42, page 165.

⁵³ Definition 5.43, page 165.

⁵⁴ Proposition 5.44, page 166.

⁵⁵ Theorem 5.46, page 167.

We found that any finitely generated abelian group can be expressed in a canonical form, called the **invariant factor decomposition**, which consists of a direct sum of finite and infinite cyclic groups.^{56,57} In particular, this means that any finite abelian group can be expressed as a direct sum of finite cyclic groups, and any torsion-free abelian group must be isomorphic to a free abelian group $FA_r = \mathbb{Z}^r$.

⁵⁶ Definition 5.47, page 169.

⁵⁷ Theorem 5.48, page 169.

In practice, we can take any presentation for a finitely generated abelian group, write down the corresponding coefficient matrix and use a finite sequence of elementary row and column operations⁵⁸ to convert it into **Smith normal form**.⁵⁹ Having done this, we can

⁵⁸ Definition 5.57, page 175.

⁵⁹ Proposition 5.58, page 175.

⁶⁰ Algorithm 5.59, page 175.

⁶¹ Definition 5.62, page 178.

⁶² Theorem 5.63, page 178.

⁶³ Definition 5.66, page 180.

⁶⁴ Definition 5.71, page 185.

⁶⁵ Definition 5.73, page 186.

⁶⁶ Definition 5.76, page 188.

⁶⁷ Proposition 5.77, page 188.

⁶⁸ Theorem 5.11, page 143.

⁶⁹ Theorem 5.46, page 167.

immediately read off the invariant factor decomposition.⁶⁰ The process is very similar to the usual linear algebraic method for solving systems of simultaneous linear equations.

Any finitely generated abelian group can also be canonically decomposed in a different way to give the **primary decomposition**.^{61,62}

Presentations are often a useful way of describing a group in a compact form, but as noted earlier their usefulness is limited by certain technical constraints (such as the word problem). We can't, for example, immediately tell whether a given presentation determines an infinite group, or if finite, what its order is. In the case of a finite group, however, the **Todd–Coxeter algorithm** can give us the order of the group, a permutation representation and a Cayley graph⁶³. If applied to an infinite group, the algorithm won't terminate. (This presents us with a potential logistical problem: there is no way of telling whether the algorithm hasn't terminated because the group is infinite, or just that it hasn't terminated *yet* because the group is finite but very large.) A slightly modified form of the Todd–Coxeter algorithm can be used to find the index $|G : H|$ of a subgroup $H \leq G$ and also a **transversal**:⁶⁴ a set of group elements, one from each right coset of H in G .

The transversals obtained by the Todd–Coxeter algorithm have a special property: they're **prefix-closed**. We call transversals satisfying this property **Schreier transversals**,⁶⁵ and it turns out that with the aid of the **shortlex** well-ordering relation⁶⁶ we can show that any subgroup H of a free group F has a transversal of this type.⁶⁷

We can use this machinery to prove the Nielsen–Schreier Theorem,⁶⁸ which states that a subgroup H of a free group F is itself free; moreover if the rank $r = \text{rank}(F)$ and the index $s = |F : H|$ are both finite, then $\text{rank}(H)$ is also finite and equal to $(r-1)s + 1$. It is a nonabelian analogue of Dedekind's Theorem,⁶⁹ which makes a similar statement about subgroups of free abelian groups.

The techniques used in the proof of the Nielsen–Schreier Theorem also yield a general method for calculating presentations of a subgroup H of a given finitely presented group G . We use the Todd–Coxeter algorithm to obtain a Schreier transversal U , which we then use to construct generators and relators for H .

References and further reading

- D L Johnson, *Presentations of Groups*, second edition, London Mathematical Society Student Texts 15,

Cambridge University Press (1997)

A good textbook on combinatorial group theory, aimed primarily at graduate students and final-year undergraduates, which goes into more detail and covers several additional topics.

- S Roberts, *King of Infinite Space*, Profile Books (2007)
An excellent biography of HSM Coxeter, containing many details of his life and career, as well as his numerous contributions to algebra and geometry.
- JE Humphreys, *Reflection Groups and Coxeter Groups*, Cambridge Studies in Advanced Mathematics 29, Cambridge University Press (1990)
A readable textbook on Coxeter groups, suitable for graduate students and final-year undergraduates.

Exercises

- 5.1 As in Exercise 1.23, let $S = \mathbb{R} \setminus \{-1, 0, 1\}$, let $f: S \rightarrow S$ such that $f(x) = \frac{1+x}{1-x}$, and denote by f^n the n -fold composite of f . Furthermore, let $g: S \rightarrow S$ with $g(x) = -x$. Show that $\langle f, g \rangle \cong D_n$ for some n .
- 5.2 Write down a presentation for the group $\mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}$.
- 5.3 Show that $|Dic_n| = 4n$ by a similar argument to that in the proof of Proposition 5.30.
- 5.4 Use a sequence of Tietze transformations to transform the presentation

$$\langle x, y : xyx = yxy \rangle$$

into the presentation

$$\langle a, b : a^3 = b^2 \rangle.$$

- 5.5 List the invariant factor decompositions of all abelian groups of order 60.
- 5.6 List all the words of length 3 in the free group F_2 and put them in shortlex order.
- 5.7 List the words of length at most 2 in the free group $F_3 = \langle x, y, z : \rangle$ and put them in shortlex order.
- 5.8 Let $<$ be the shortlex ordering on some free group F with basis X . Suppose that $v, w \in F$ and $x \in X^\pm$ such that $v < w$. Show that $vx < wx$ and $xv < xw$.
- 5.9 Repeat Example 5.80 using the other Schreier transversal $U = \{e, x, y, x^2, xy, x^2y\}$ from Example 5.79.

6 Actions

The Earth is full of anger,
 The seas are dark with wrath,
 The Nations in their harness
 Go up against our path:
 Ere yet we loose the legions –
 Ere yet we draw the blade,
 Jehovah of the Thunders,
 Lord God of Battles, aid!
 — Rudyard Kipling (1865–1936),
Hymn Before Action (1896)

MANY OF THE GROUPS WE’VE STUDIED SO FAR have consisted of operations acting on a set. The dihedral and other symmetry groups, for example, are concerned with certain transformations which map a particular geometric shape to itself. Matrix groups such as $GL_n(\mathbb{R})$ or SU_n can be regarded as groups of invertible linear transformations defined on some real or complex vector space. And permutation groups such as S_n and A_n are defined precisely in terms of particular operations on some set of objects.

In this chapter, we will investigate this approach in more detail, formulating a rigorous definition of what it means for a group to *act* on a set, studying what it means for an element of a group to fix one or more elements of the chosen set in place, and examining some powerful applications to combinatorics.

6.1 Symmetries and transformations

RECALL THAT IN CHAPTER 1 we met a class of groups consisting of symmetry operations defined on some geometric object.

The dihedral groups D_n , for example, consist of rotations and reflections of a regular n -sided polygon. More precisely, we define a subset P_n of $\mathbb{R}^2$ consisting of the regular polygon with vertices at the points $(\cos \frac{(2k-1)\pi}{n}, \sin \frac{(2k-1)\pi}{n})$ for $1 \leq k \leq n$.

The reflection m_1 determines a particular bijection $P_n \rightarrow P_n$ which leaves every point along the line $y = x \tan \frac{\pi}{n}$ fixed where it is, and swaps every other point with its mirror image in that line. Similarly, the rotation r fixes the origin and cycles all other points among themselves in groups of n , and thus also determines a bijection $P_n \rightarrow P_n$.

So, what we have here is a well-defined way of combining an element of D_n together with a point in P_n to get another point in P_n ; that is, a function $D_n \times P_n \rightarrow P_n$. (Another way of looking at it is that we have a

The chief forms of beauty are order and symmetry and definiteness, which the mathematical sciences demonstrate in a special degree.

— Aristotle (384–322 BC),
Metaphysics XIII:3

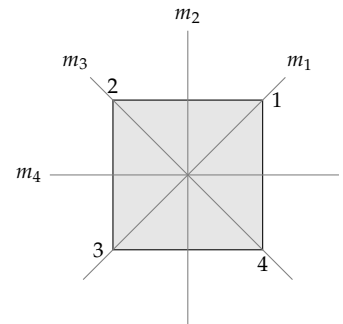


Figure 6.1: Axes of symmetry of the square P_4

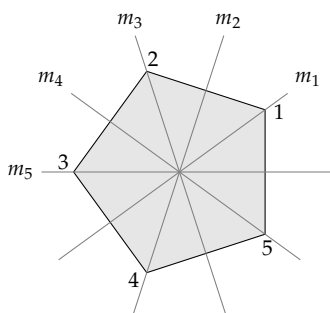


Figure 6.2: Axes of symmetry of the pentagon P_5

well-defined way of interpreting elements of D_n as bijections $P_n \rightarrow P_n$, and we'll compare these two viewpoints in a little while.) In particular, the identity element e leaves every point where it is. Also, for any two elements $g, h \in D_n$, applying h to a point $x \in P_n$ and then applying g to the result has the same effect as if we'd applied the transformation corresponding to $g * h$ to x .

Writing this down formally, we have a set P_n together with a function $\alpha: D_n \times P_n \rightarrow P_n$ such that

$$\alpha(e, x) = x,$$

$$\alpha(g, \alpha(h, x)) = \alpha(g * h, x).$$

Now let A be an invertible $n \times n$ matrix with real entries. We know from linear algebra that such a matrix can be regarded as representing a linear transformation $f_A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ (relative to some basis for $\mathbb{R}^n$). The invertibility of A ensures this linear transformation is bijective.

In particular, the identity matrix leaves every point in $\mathbb{R}^n$ fixed, and we know that $(AB)\mathbf{v} = A(B\mathbf{v})$ for any two such invertible $n \times n$ matrices A and B , and any vector $\mathbf{v} \in \mathbb{R}^n$. What we have, then, is a function

$$GL_n(\mathbb{R}) \times \mathbb{R}^n \rightarrow \mathbb{R}^n.$$

(The of this function depend on our choice of basis for $\mathbb{R}^n$.)

¹ Definition 1.51, page 24.

As a third example, recall that the symmetric group S_n was defined¹ as the group of permutations acting on a set $X_n = \{1, \dots, n\}$. Again, we have an identity permutation ι which leaves X_n unchanged, and for any two permutations $\sigma, \tau \in S_n$ and element $x \in X_n$ we have $(\sigma\tau)(x) = \sigma(\tau(x))$. This, therefore, also determines a function $S_n \times X_n \rightarrow X_n$.

With this in mind, we present the following definition:

Definition 6.1 Let G be a group, and let X be a set. An **action** of G on X , or a **G -action** on X is a function

$$\alpha: G \times X \rightarrow X$$

satisfying the properties

$$\alpha(e, x) = x, \tag{6.1}$$

$$\alpha(g, \alpha(h, x)) = \alpha(g * h, x) \tag{6.2}$$

for all $x \in X$ and $g, h \in G$. The set X is then said to be a **G -set**.

² Definition 1.1, page 2.

As with the binary operations we met early in Chapter 1,² in many cases it will be simpler and clearer if we adopt a slightly different notation. Instead of writing $\alpha(g, x)$, then, we'll often write $g \cdot x$ instead. In this form, conditions (6.1) and (6.2) become much clearer:

$$e \cdot x = x, \tag{6.3}$$

$$g \cdot (h \cdot x) = (g * h) \cdot x. \tag{6.4}$$

The first says that the identity e must act as the identity operation on X , while the second is an associativity condition.

What we've just defined here is more properly called a **left action** of G on X . There is a corresponding notion of a **right action**, which is a function $\beta: X \times G \rightarrow X$ satisfying

$$\beta(x, e) = x, \quad (6.5)$$

$$\beta(\beta(x, g), h) = \beta(x, g * h), \quad (6.6)$$

or, alternatively,

$$x \cdot e = x, \quad (6.7)$$

$$(x \cdot g) \cdot h = x \cdot (g * h), \quad (6.8)$$

for all $x \in X$ and $g \in G$.

It so happens that there is a bijective mapping between left and right G -actions on a given set X , in the sense that for any left action $\alpha: G \times X \rightarrow X$ there is a unique right action $\beta: X \times G \rightarrow X$ and vice versa.

This correspondence doesn't work in quite the way we might expect, however. The tempting but naïve approach is to take a left action $\alpha: G \times X \rightarrow X$ and define a right action $\beta: X \times G \rightarrow X$ by setting $\beta(x, g) = \alpha(g, x)$ for all $g \in G$ and $x \in X$.

But this doesn't quite work. It certainly satisfies condition (6.7) since

$$\beta(x, e) = \alpha(e, x) = x.$$

Condition (6.8), however, fails in general when G is nonabelian:

$$\beta(\beta(x, g), h) = \alpha(h, \alpha(g, x)) = \alpha(h * g, x) = \beta(x, h * g) \neq \beta(x, g * h).$$

We can make this work, however, if instead of defining $\beta(x, g) = \alpha(g, x)$ we define $\beta(x, g) = \alpha(g^{-1}, x)$ for all $g \in G$ and $x \in X$. Here,

$$\beta(x, e) = \alpha(e^{-1}, x) = \alpha(e, x) = x$$

as before, but

$$\begin{aligned} \beta(\beta(x, g), h) &= \alpha(h^{-1}, \alpha(g^{-1}, x)) = \\ &= \alpha(h^{-1} * g^{-1}, x) = \alpha((g * h)^{-1}, x) = \beta(x, g * h). \end{aligned}$$

In the three motivating examples earlier, we interpreted the action of D_n on P_n , the action of $GL_n(\mathbb{R})$ on $\mathbb{R}^n$ and the action of S_n on X_n as functions of the form $G \times X \rightarrow X$ satisfying the properties in Definition 6.1. But in all three cases we noted that the action of a particular group element on the set determined a bijective symmetry map, invertible linear transformation or a permutation of that set. So we could also regard those group actions as ways of assigning a permutation of X to each element of the group G . This is true in general, as the following proposition shows.

Proposition 6.2 Let $\alpha: G \times X \rightarrow X$, defined by $(g, x) \mapsto g \cdot x$ for $g \in G$ and $x \in X$, be a left action of a group G on a set X .

Then α determines a unique homomorphism $\phi: G \rightarrow \text{Sym}(X)$, given by mapping an element $g \in G$ to the function $f_g: X \rightarrow X$ where $f_g(x) = \alpha(g, x) = g \cdot x$ for all $x \in X$.

Proof Property (6.3) tells us that $f_e(x) = e \cdot x = x$ for all $x \in X$, so f_e is the identity map id_X .

Property (6.4) tells us that

$$f_e(x) = f_{g * g^{-1}}(x) = (g * g^{-1}) \cdot x = g \cdot (g^{-1} \cdot x) = f_g(f_{g^{-1}}(x))$$

and also that

$$f_e(x) = f_{g^{-1} * g}(x) = (g^{-1} * g) \cdot x = g^{-1} \cdot (g \cdot x) = f_{g^{-1}}(f_g(x)).$$

This tells us that $f_g \circ f_{g^{-1}} = f_e = f_{g^{-1}} \circ f_g$, so f_g must be a bijection for all $g \in G$, and hence $f_g \in \text{Sym}(X)$. So the function $\phi: G \rightarrow \text{Sym}(X)$ given by $g \mapsto f_g$ is certainly well-defined, and we just need to show that it's also a homomorphism.

To see this, let $g, h \in G$. Then

$$\begin{aligned} \phi(g * h)(x) &= f_{g * h}(x) = (g * h) \cdot x = g \cdot (h \cdot x) = f_g(f_h(x)) \\ &= (f_g \circ f_h)(x) = (\phi(g) \circ \phi(h))(x) \end{aligned}$$

as required. $\square$

Similarly, any homomorphism $\phi: G \rightarrow \text{Sym}(X)$ determines a unique left action of G on X , and since left G -actions on X are in bijective correspondence with right G -actions on X , any such homomorphism ϕ determines a unique right action too.

So G -actions on a set X are essentially the same as homomorphisms $G \rightarrow \text{Sym}(X)$. We studied homomorphisms quite extensively in Chapter 4, so by applying some of the things we learned, we should be able to get some important insights into how group actions work.

Example 6.3 By examining how the elements of D_4 permute the vertices of the square P_4 , as shown in Figure 6.3, we can obtain an action of D_4 on the set $X_4 = \{1, 2, 3, 4\}$. This determines a homomorphism $\phi: D_4 \rightarrow S_4$ such that

$$\begin{array}{ll} e \mapsto \iota, & r \mapsto (1234), \\ r^2 \mapsto (13)(24), & r^3 \mapsto (1432), \\ m_1 \mapsto (24), & m_2 \mapsto (12)(34), \\ m_3 \mapsto (13), & m_4 \mapsto (14)(23). \end{array}$$

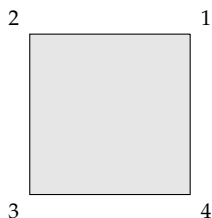


Figure 6.3: Vertices of the square P_4

The dihedral group D_n acts by rotation and reflection operations on a regular n -sided polygon P_n , but as we noted at the beginning of

Section 1.4 we can also view it as acting on the subset consisting of just the n vertices of P_n . We'll come back to this idea of a group acting on a subset in the next section, when we consider orbits and stabilisers. But by observing just the behaviour of the n vertices of P_n under the action of D_n , we can define a homomorphism $\phi: D_n \rightarrow S_n$.³

In particular, it's worth thinking about what happens when the homomorphism $\phi: G \rightarrow \text{Sym}(X)$ is injective. If this is the case then any two distinct elements $g_1, g_2 \in G$ map to different permutations in $\text{Sym}(X)$. This means that there exists at least one $x \in X$ for which $g_1 \cdot x \neq g_2 \cdot x$. In other words, each $g \in G$ acts differently on the elements of X .

If ϕ isn't injective, then we'll be able to find two distinct elements $g_1, g_2 \in G$ which act in the same way on all elements of the set X . Specifically, there will be a nontrivial element $g \in G$ which acts trivially on X ; that is, $g \cdot x = x$ for all $x \in X$.

If $\phi: G \rightarrow \text{Sym}(X)$ is injective, each element of G permutes the elements of X in a different way. We give a special name to such actions:

Definition 6.4 Let a group G act on a set X in such a way that the corresponding homomorphism $\phi: G \rightarrow \text{Sym}(X)$ is injective. Then we say that this action is **faithful**.

Equivalently, a G -action on X is faithful if, for any distinct elements $g_1, g_2 \in G$, there exists at least one $x \in X$ for which $g_1 \cdot x \neq g_2 \cdot x$.

Proposition 4.35 says that injective homomorphisms are exactly those that have trivial kernel. So in the context of group actions the kernel of the homomorphism $\phi: G \rightarrow \text{Sym}(X)$ consists of those elements of G that are mapped to the identity permutation $\iota \in \text{Sym}(X)$ and hence act trivially on X . In other words, a faithful action is one for which only the identity $e \in G$ acts trivially on the set X .

Example 6.5 Perhaps the most obvious group action is that of the full symmetric group $\text{Sym}(X)$ on a set X . This action corresponds to the identity homomorphism $\text{id}: \text{Sym}(X) \rightarrow \text{Sym}(X)$, and since this is injective the action is faithful.

The next most obvious group action is the one where nothing happens:

Example 6.6 For any group G and any set X we can define the **trivial action** of G on X by setting $g \cdot x = x$ for all $g \in G$ and $x \in X$. This certainly satisfies conditions (6.3) and (6.4) and is hence a valid example of a group action. From the alternative point of view of a homomorphism $G \rightarrow \text{Sym}(X)$ this is exactly the homomorphism that maps every element $g \in G$ to the identity permutation $\iota \in \text{Sym}(X)$. This action is not faithful in general, except when the group G is trivial.

³ To be really precise about this, we first have to number the vertices by defining a bijection between the n vertices and the set $\{1, 2, \dots, n\}$. There are $n!$ ways of doing this, and the resulting homomorphism $\phi: D_n \rightarrow S_n$ will depend on exactly which choice we make, but not to any really important degree: any two such homomorphisms ϕ_1 and ϕ_2 will differ only by composition with some automorphism $f: S_n \rightarrow S_n$, in the sense that $\phi_2 = f \circ \phi_1$.

Suppose G is a subgroup of the symmetric group $\text{Sym}(X)$ for some set X . Then each element of G is still a permutation of X , so we can construct an action of G on X by using the inclusion homomorphism $i: G \hookrightarrow \text{Sym}(X)$. What does the corresponding action look like? Well, it's exactly what we'd expect it to be: an element $g \in G$ is a permutation of X ; that is, a bijection $g: X \rightarrow X$. So we can define $g \cdot x = g(x)$ for all $x \in X$ and get a G -action on X that satisfies the conditions (6.3) and (6.4).

The inclusion homomorphism $i: G \hookrightarrow \text{Sym}(X)$ is injective and hence this action is faithful.

Example 6.7 The alternating group A_n is a subgroup of the finite symmetric group S_n , and so there is an inclusion homomorphism $i: A_n \hookrightarrow S_n$ via which A_n acts on the finite set $X_n = \{1, \dots, n\}$ in the obvious way.

We can take this viewpoint a little further. Suppose that a group G acts on some set X , and suppose also that H is some subgroup of G . Then we can define an action of H on X in a straightforward way: every element of H is also an element of G after all, so for $h \in H$ and $x \in X$ we can just define $h \cdot x$ to be whatever it is in the action of G on X .

Another way of looking at this is to define the homomorphism $\psi: H \rightarrow \text{Sym}(X)$ by composing the inclusion homomorphism $i: H \hookrightarrow G$ with the action homomorphism $\phi: G \rightarrow \text{Sym}(X)$.

Example 6.8 A concrete example of this is given by the inclusion of the special orthogonal group $SO_n(\mathbb{R})$ in the general linear group $GL_n(\mathbb{R})$. Once we've chosen a basis for $\mathbb{R}^n$, that gives us an action of $GL_n(\mathbb{R})$ on $\mathbb{R}^n$: a well-defined method of combining an invertible $n \times n$ real matrix with an n -component real vector to get another n -component real vector, in such a way that the usual $\mathbb{R}$ -linearity conditions are satisfied. Whatever other properties (unit determinant, orthogonality) elements of $SO_n(\mathbb{R})$ might have, they are still $n \times n$ invertible real matrices, so we can define an $SO_n(\mathbb{R})$ -action on $\mathbb{R}^n$ by reusing the $GL_n(\mathbb{R})$ action we already had.

⁴ Theorem 2.13, page 47.

By Cayley's Theorem⁴, any group G can be regarded as a subgroup of the symmetric group $\text{Sym}(G)$ of permutations of its underlying set.

More precisely, we construct an isomorphism $f: G \rightarrow H < \text{Sym}(G)$ by mapping each element $g \in G$ to the bijection $\lambda_g: G \rightarrow G$ defined by $\lambda_g(h) = g * h$ for all $h \in G$. This yields an isomorphism

$$f: G \rightarrow H = \{\lambda_g : g \in G\} \leq \text{Sym}(G).$$

We can compose this isomorphism f with the inclusion homomorphism $i: H \hookrightarrow \text{Sym}(G)$ to get the required action homomorphism $\phi = i \circ f: G \rightarrow \text{Sym}(G)$.

We can do this for any group G , finite or infinite, to obtain an action of G on its own underlying set:

Definition 6.9 For any group G we can define the **left regular action** of G on itself by defining

$$\alpha(g, h) = g \cdot h = g * h$$

for all $g, h \in G$. This action is faithful: $g \cdot h = h$ if and only if $g = e$, by the left cancellation laws (1.4).

Similarly, we can define the **right regular action** of G on itself by

$$\beta(h, g) = h \cdot g = h * g$$

for all $g, h \in G$. This action is also faithful.

We can use this process to write down another D_4 -action:

Example 6.10 Applying Cayley's Theorem to the dihedral group D_4 (see Table 6.1) gives us an isomorphism $f: D_4 \cong G \leq \text{Sym}(D_4) \cong S_8$, such that

$$\begin{aligned} e &\mapsto \iota, & r &\mapsto (e r r^2 r^3)(m_1 m_4 m_3 m_2), \\ r^2 &\mapsto (e r^2)(r r^3)(m_1 m_3)(m_2 m_4), & r^3 &\mapsto (e r^3 r^2 r)(m_1 m_2 m_3 m_4), \\ m_1 &\mapsto (e m_1)(r m_2)(r^2 m_3)(r^3 m_4), & m_2 &\mapsto (e m_2)(r m_3)(r^2 m_4)(r^3 m_1), \\ m_3 &\mapsto (e m_3)(r m_4)(r^2 m_1)(r^3 m_2), & m_4 &\mapsto (e m_4)(r m_1)(r^2 m_2)(r^3 m_3). \end{aligned}$$

Composing this isomorphism f with the inclusion homomorphism $i: G \hookrightarrow \text{Sym}(D_4)$ gives us an action of D_4 on its underlying set $D_4 = \{e, r, r^2, r^3, m_1, m_2, m_3, m_4\}$. This action is faithful: f is an isomorphism and i is injective so their composite $i \circ f: D_4 \rightarrow \text{Sym}(D_4)$ must be injective too.

	e	r	r^2	r^3	m_1	m_2	m_3	m_4
e	e	r	r^2	r^3	m_1	m_2	m_3	m_4
r	r	r^2	r^3	e	m_4	m_1	m_2	m_3
r^2	r^2	r^3	e	r	m_3	m_4	m_1	m_2
r^3	r^3	e	r	r^2	m_2	m_3	m_4	m_1
m_1	m_1	m_2	m_3	m_4	e	r	r^2	r^3
m_2	m_2	m_3	m_4	m_1	r^3	e	r	r^2
m_3	m_3	m_4	m_1	m_2	r^2	r^3	e	r
m_4	m_4	m_1	m_2	m_3	r	r^2	r^3	e

Table 6.1: The multiplication table for the dihedral group D_4

Cayley's Theorem gives a faithful action of a group on its underlying set. The group conjugation operation from Chapter 3 also satisfies (6.3) and (6.4), and yields another important action of a group on itself.

Definition 6.11 Let G be a (finite or infinite) group. Then the **conjugation action** of G on itself is defined by

$$h^g = g \cdot h = g * h * g^{-1}$$

for all $g, h \in G$. This action is not in general faithful.

Let's look at a simple example first.

Example 6.12 The conjugation action of $\mathbb{Z}_3$ on itself (see Table 6.2) yields the homomorphism $\mathbb{Z}_3 \rightarrow \text{Sym}(\mathbb{Z}_3) \cong S_3$ given by

$$0 \mapsto \iota, \quad 1 \mapsto \iota, \quad 2 \mapsto \iota.$$

This is the trivial action on $\mathbb{Z}_3$, and is obviously not faithful.

	0	1	2
0	0	0	0
1	1	1	1
2	2	2	2

Table 6.2: Conjugation in $\mathbb{Z}_3$

In general, any abelian group acts trivially on itself by conjugation:

Proposition 6.13 *Let G be an abelian group. Then the conjugation action on G is trivial.*

Proof Let $g, h \in G$. Then $h^g = g * h * g^{-1} = g * g^{-1} * h = e * h = h$. Therefore the conjugation action is trivial. $\square$

Example 6.14 Writing the corresponding permutations in $\text{Sym}(D_4)$ in cycle notation, as in Example 6.10, we obtain

	e	r	r^2	r^3	m_1	m_2	m_3	m_4
e	e	e	e	e	e	e	e	e
r	r	r	r	r	r^3	r^3	r^3	r^3
r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2	r^2
r^3	r^3	r^3	r^3	r^3	r	r	r	r
m_1	m_1	m_3	m_1	m_3	m_1	m_3	m_1	m_3
m_2	m_2	m_4	m_2	m_4	m_4	m_2	m_4	m_2
m_3	m_3	m_1	m_3	m_1	m_3	m_1	m_3	m_1
m_4	m_4	m_2	m_4	m_2	m_2	m_4	m_2	m_4

Table 6.3: Conjugation in D_4

$$\begin{aligned}
 e &\mapsto \iota, & r &\mapsto (m_1 m_3)(m_2 m_4), \\
 r^2 &\mapsto \iota, & r^3 &\mapsto (m_1 m_3)(m_2 m_4), \\
 m_1 &\mapsto (r r^3)(m_2 m_4), & m_2 &\mapsto (r r^3)(m_1 m_3), \\
 m_3 &\mapsto (r r^3)(m_2 m_4), & m_4 &\mapsto (r r^3)(m_1 m_3).
 \end{aligned}$$

This action is clearly not faithful, since the kernel of the homomorphism $D_4 \rightarrow \text{Sym}(D_4)$ is nontrivial: it consists of e and r^2 .

So far we've met a number of different group actions, some of which are faithful and some of which aren't. Proposition 4.29 says that kernels of group homomorphisms correspond exactly to normal subgroups. So another way to look at this is that a nontrivial unfaithful G -action on a set X determines a nontrivial normal subgroup of G .

This, together with the First Isomorphism Theorem⁵ gives us a method for turning any unfaithful group action into a faithful one.

An unfaithful G -action on a set X gives a noninjective homomorphism $\phi: G \rightarrow \text{Sym}(X)$. Factoring out by the kernel $K = \ker(\phi)$ of this homomorphism yields an isomorphism $G/K \rightarrow \text{im}(\phi)$, and composing this with the inclusion homomorphism $i: \text{im}(\phi) \hookrightarrow \text{Sym}(X)$ gives us an injective homomorphism $\psi: G/K \rightarrow \text{Sym}(X)$. This homomorphism ψ determines a faithful action of G/K on X .

Example 6.15 The unfaithful conjugation D_4 -action from Example 6.14 gives a noninjective homomorphism $\phi: D_4 \rightarrow \text{Sym}(D_4)$ with kernel $K = \{e, r^2\}$. From Example 3.30 we know that $D_4/K \cong V$, the Klein group, so by factoring out by K (which happens to be the centre $Z(D_4)$) we get a faithful action of V on the underlying set of D_4 given by the following injective homomorphism $\psi: V \rightarrow \text{Sym}(D_4)$:

$$\begin{aligned}
 e &\mapsto \iota, & a &\mapsto (m_1 m_3)(m_2 m_4), \\
 b &\mapsto (r r^3)(m_2 m_4), & c &\mapsto (r r^3)(m_1 m_3).
 \end{aligned}$$

Also, this relationship between group actions and homomorphisms and normal subgroups means that if G is a simple group, then the only unfaithful G -action is the trivial one.

⁵ Theorem 4.40, page 117.

6.2 Orbits and stabilisers

By PROPOSITION 1.54, any finite permutation can be decomposed as a product of disjoint cyclic permutations. For example, the permutation $\tau \in S_5$ given by $\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 5 & 1 & 3 \end{bmatrix}$ can be written as the product $(1\ 2\ 4)(3\ 5)$. When this permutation acts on the set $X_5 = \{1, 2, 3, 4, 5\}$ it cycles 1, 2 and 4 amongst themselves, and swaps 3 and 5. We can use this to define an action of the cyclic subgroup

$$\langle \tau \rangle = \{ \iota, (1\ 2\ 4)(3\ 5), (1\ 4\ 2), (3\ 5), (1\ 2\ 4), (1\ 4\ 2)(3\ 5) \} \leq S_5$$

on X_5 in the obvious way. Looking at how the six elements of $\langle \tau \rangle$ permute the elements of X_5 , we can see that all of them permute 1, 2 and 4 amongst themselves, and also permute 3 and 5 amongst themselves. No element of $\langle \tau \rangle$, though, maps an element of the subset $\{1, 2, 4\}$ to an element of the subset $\{3, 5\}$, or vice versa. (This is obvious, because since $\tau = (1\ 2\ 4)(3\ 5)$ doesn't do this, neither can any power of τ .) So, the action of $\langle \tau \rangle$ on X_5 splits (or **partitions**) X_5 as a union of two disjoint subsets: $X_5 = \{1, 2, 4\} \cup \{3, 5\}$.

Where there's a partition, there's an equivalence relation: we can say two elements $x, y \in S = \bigcup_i S_i$ of a partitioned set S are related, written $x \sim y$, if they both lie in the same subset $S_i \subseteq S$. This relation is reflexive (x must obviously lie in the same subset as itself), it's symmetric (if x lies in the same subset as y then y must also be in the same subset as x) and it's transitive (if x is in the same subset as y and y is in the same subset as z then x must be in the same subset as z).

An arbitrary partition needn't correspond to a particularly interesting equivalence relation, but since this one is determined by the action of $\langle \tau \rangle$ on X_5 , we might reasonably expect the corresponding equivalence relation to give us some insight into exactly how $\langle \tau \rangle$ acts on X_5 .

The key is the observation a few paragraphs ago: the elements of $\langle \tau \rangle$ permute the subsets $\{1, 2, 4\}$ and $\{3, 5\}$ amongst themselves, but they don't map any elements from one subset to the other. More formally, for two elements $x, y \in X_5$, we say that $x \sim y$ if and only if there exists a permutation $\sigma \in \langle \tau \rangle$ such that $\sigma x = y$.

We need a name for the equivalence classes determined by this action.

Definition 6.16 Let G be a group, acting on a set X . The **orbit** of an element $x \in X$ is the subset

$$\text{Orb}_G(x) = \{y \in X : g \cdot x = y \text{ for some } g \in G\}.$$

So, in the example above,

$$\text{Orb}_{\langle \tau \rangle}(1) = \text{Orb}_{\langle \tau \rangle}(2) = \text{Orb}_{\langle \tau \rangle}(4) = \{1, 2, 4\}$$

$$\text{and } \text{Orb}_{\langle \tau \rangle}(3) = \text{Orb}_{\langle \tau \rangle}(5) = \{3, 5\}.$$

[The Earth] alone remains immoveable, whilst all things revolve around it, being connected with every other part, while they all rest upon it.

— Pliny the Elder (23–79AD),
Natural History II:4

If M 's a complete metric space, and nonempty, it's always the case, that if f 's a contraction, then under its action, exactly one point stays in place.

— Anonymous, *The Contraction Mapping Limerick*

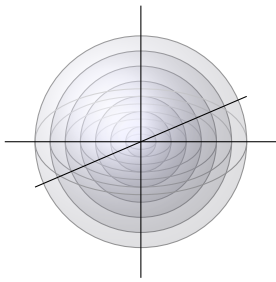


Figure 6.4: Orbits of $SO_3(\mathbb{R})$ as concentric spheres

The next example involves an infinite group acting on an infinite set.

Example 6.17 The group $SO_3(\mathbb{R})$ consists of all real, orthogonal 3×3 matrices with determinant 1. Geometrically, these correspond to rotations in $\mathbb{R}^3$ around some axis passing through the origin.

What are the orbits of this $SO_3(\mathbb{R})$ -action on $\mathbb{R}^3$? Well, given any two position vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^3$, if $\|\mathbf{v}\| = \|\mathbf{w}\|$ then there will be a matrix $A \in SO_3(\mathbb{R})$ such that $A\mathbf{v} = \mathbf{w}$. But if $\|\mathbf{v}\| \neq \|\mathbf{w}\|$ then no such special orthogonal matrix will exist. So $\mathbf{v} \sim \mathbf{w}$ if and only if $\|\mathbf{v}\| = \|\mathbf{w}\|$, which means that the orbit $\text{Orb}_{SO_3(\mathbb{R})}(\mathbf{v})$ must consist of all vectors $\mathbf{w} \in \mathbb{R}^3$ such that $\|\mathbf{v}\| = \|\mathbf{w}\|$. Geometrically, these position vectors determine a sphere of radius $\|\mathbf{v}\|$ centred on the origin. So the action of $SO_3(\mathbb{R})$ partitions $\mathbb{R}^3$ into an infinite collection of concentric spheres and the origin $\{\mathbf{0}\} = \text{Orb}_{SO_3(\mathbb{R})}(\mathbf{0})$.

The following example is of a finite group, the dihedral group D_3 , acting in a familiar way on an infinite set of points in the plane.

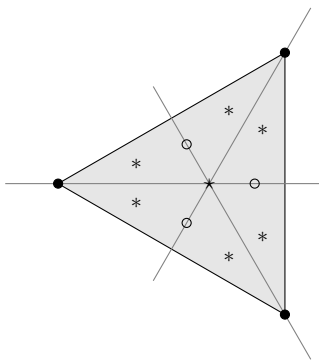


Figure 6.5: Axes of symmetry of the equilateral triangle P_3 , and some typical orbits of the D_3 -action

Example 6.18 Let P_3 be the equilateral triangle with vertices at the points $(-1, 0)$, $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ and $(\frac{1}{2}, -\frac{\sqrt{3}}{2})$. Then the dihedral group D_3 acts on this subset of the plane $\mathbb{R}^2$ in the usual way, and this action has uncountably many orbits. These orbits, however, are of three different types:

The origin $(0, 0)$ comprises an orbit on its own: it's left fixed by every element in D_6 .

The three vertices form an orbit: as remarked at the beginning of Section 1.4 they are permuted among themselves by the elements of D_6 . More generally, any point on one of the triangle's axes of symmetry, apart from the origin itself, forms an orbit with two other points.

Finally, any point which doesn't lie on an axis of symmetry forms an orbit with five other points. See Figure 6.5 for an illustration of some typical orbits.

One of the standard group actions we met in the last section was the trivial action. Its orbits have a particularly simple form:

Example 6.19 Let a group G act trivially on some set X , in the sense that $g \cdot x = x$ for all $g \in G$ and $x \in X$. Then for some arbitrary $x \in X$, the orbit $\text{Orb}_G(x) = \{g \cdot x : g \in G\} = \{x\}$. So the trivial action partitions a set X into singleton subsets consisting of the individual elements of X .

Another important action is the conjugation action of a group G acting on itself.⁶ We met the orbits of this action in a slightly different context in Chapter 3.⁷

⁶ Definition 6.11, page 211.

⁷ Proposition 3.4, page 72.

Example 6.20 Let a group G act on its underlying set via the conjugation action. Then the orbit of an element $h \in G$ is the set

$$\text{Orb}_G(h) = \{g \cdot h : g \in G\} = \{g * h * g^{-1} : g \in G\}.$$

But this is exactly the conjugacy class of h in G . Two elements $h, k \in G$ are therefore in the same orbit if they are conjugate: if there exists some element $g \in G$ such that $k = g * h * g^{-1}$.

Sometimes, all elements of a G -set X lie in the same orbit. The action of S_n on X_n (or, more generally, of $\text{Sym}(X)$ on a set X) has this property. That is, for any two elements $x, y \in X$, there is at least one element $g \in G$ for which $g \cdot x = y$. Another concrete example is the action of $GL_n(\mathbb{R})$ on $\mathbb{R}_*^3 = \mathbb{R}^3 \setminus \{0\}$: for any nonzero vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}_*^3$ we can find an invertible $n \times n$ real matrix A such that $A\mathbf{v} = \mathbf{w}$.

Definition 6.21 An action of a group G on a set X is **transitive** if it partitions X into a single orbit. Equivalently, for any $x, y \in X$ there exists at least one element $g \in G$ such that $g \cdot x = y$.

In Section 3.A we'll extend these ideas to obtain **multiply transitive** actions, which we will use to construct the **Mathieu groups**.

Proposition 6.22 *The left regular action of a group on itself is transitive.*

Proof This is really just a consequence of the cancellation law for group multiplication.⁸ To see this, consider any two elements $h, k \in G$. For h and k to be in the same orbit, there must be an element $g \in G$ such that $g \cdot h = g * h = k$. Setting $g = k * h^{-1}$ works, so the left regular action is transitive. $\square$

⁸ Proposition 1.15, page 8.

The right regular action is transitive by a similar argument.

For the moment, though, we'll return to the D_3 -action described in Example 6.18, and look at it from a slightly different perspective. Instead of asking which points get mapped to which others by arbitrary elements of the group, we ask what group elements leave a given point of the triangle P_3 unchanged.

The central point of the triangle P_3 is unmoved by any element of D_3 , but this isn't the case for any other point. The vertex 1 of P_3 is obviously left fixed in place by the identity e , and also by the reflection m_1 in the axis which passes through that vertex, but any other element of D_3 maps it to one of the other two vertices. This is also true for any non-central point which lies on one of the three axes of symmetry: it's fixed in place by e and the reflection operation corresponding to its axis, but the other four elements of D_3 move it to one of the other two points in the same orbit. And if we choose a point of P_3 that doesn't lie on an axis of symmetry, then it's even worse: only the identity e leaves the point where it is.

So, the triangle's central point is fixed by all of D_3 , any other point on an axis of symmetry is fixed by just a subset of the form $\{e, m_1\}$, and a point which isn't on an axis of symmetry is fixed only by the identity element e . We give subsets of this type a special name:

Definition 6.23 Let a group G act on a set X , and suppose x is some element of X . The **stabiliser** of x in G is the set

$$\text{Stab}_G(x) = \{g \in G : g \cdot x = x\}$$

of elements in G which act trivially on x .

In the D_3 example, all the stabilisers were subgroups of D_3 rather than just subsets, and it's reasonable to ask whether this is true in general: if a group G acts on a set X , then is the stabiliser $\text{Stab}_G(x)$ of a given point $x \in X$ always a subgroup of G ? It so happens that it is.

Proposition 6.24 *Let a group G act on a set X . Then for any $x \in X$, the set $\text{Stab}_G(x)$ is a subgroup of G . Furthermore, $\bigcap_{x \in X} \text{Stab}_G(x)$ is the kernel of the action homomorphism $\phi: G \rightarrow \text{Sym}(X)$.*

Proof To show that $\text{Stab}_G(x)$ is a subgroup of G , we need to check that for any two elements $g_1, g_2 \in \text{Stab}_G(x)$ the product $g_1 * g_2$ is also in $\text{Stab}_G(x)$, and that if $g \in \text{Stab}_G(x)$ then $g^{-1} \in \text{Stab}_G(x)$ as well.

To see the first of these, observe that

$$(g_1 * g_2) \cdot x = g_1 \cdot (g_2 \cdot x) = g_1 \cdot x = x$$

since both g_1 and g_2 act trivially on x . The second condition is also satisfied, since for any $g \in G$ we have

$$g^{-1} \cdot x = g^{-1} \cdot (g \cdot x) = (g^{-1} * g) \cdot x = e \cdot x = x.$$

Hence $\text{Stab}_G(x) \leq G$.

Now suppose that $g \in \bigcap_{x \in X} \text{Stab}_G(x)$. This means that $g \cdot x = x$ for all $x \in X$, which tells us that the action homomorphism ϕ maps g to the identity homomorphism ι in $\text{Sym}(X)$. Hence $\bigcap_{x \in X} \text{Stab}_G(x) \subseteq \ker(\phi)$. Conversely, suppose $g \in \ker(\phi)$, so that $\phi(g) = \iota \in \text{Sym}(X)$. Then $g \cdot x = \iota(x) = x$ for any $x \in X$, so $g \in \bigcap_{x \in X} \text{Stab}_G(x)$, and hence $\ker(\phi) \subseteq \bigcap_{x \in X} \text{Stab}_G(x)$.

Therefore $\bigcap_{x \in X} \text{Stab}_G(x) = \ker(\phi)$ as claimed. $\square$

In light of the above result, the stabiliser $\text{Stab}_G(x)$ is sometimes called the **isotropy subgroup**.

Returning yet again to the D_3 example, we notice that the orbit of the triangle P_3 's central point consists of a single element, and that its stabiliser consists of all six elements of D_3 . If we choose some other point on a symmetry axis of P_3 then we find that the orbit of this point contains three points and its stabiliser contains two elements of D_3 .

The orbit of some other point in P_3 , meanwhile, consists of six points while its stabiliser is just the identity element e .

In each of these three cases, the number of points in the orbit $\text{Orb}_{D_3}(x)$ multiplied by the number of elements in the stabiliser $\text{Stab}_{D_3}(x)$ is equal to the number of elements in the group D_3 . That is,

$$|\text{Orb}_{D_3}(x)| \times |\text{Stab}_{D_3}(x)| = 6 = |D_3|.$$

The next theorem confirms that this is true for any (finite) group acting on a (finite or infinite) set.

Theorem 6.25 (Orbit–Stabiliser Theorem) *Let G be a finite group acting on a (not necessarily finite) set X . Then*

$$|\text{Orb}_G(x)| \times |\text{Stab}_G(x)| = |G|$$

for all $x \in X$.

Proof Let y be some element of the orbit $\text{Orb}_G(x)$. Then there exists at least one element $g \in G$ such that $g \cdot x = y$. Now suppose that there exists some other $h \in G$ such that $h \cdot x = y$. Then we have $h \cdot x = g \cdot x$, and hence $(g^{-1} \cdot h) \cdot x = x$. This means that $g^{-1} \cdot h$ is in the stabiliser $\text{Stab}_G(x)$. Proposition 6.24 tells us that $\text{Stab}_G(x)$ is a subgroup of G , and hence by Proposition 2.26 it follows that if $g^{-1} \cdot h$ is in $\text{Stab}_G(x)$ then h must be in the left coset $g \text{Stab}_G(x)$.

So all group elements which map x to y lie in the same coset of $\text{Stab}_G(x)$. Conversely, if an element $k \in G$ lies in the coset $g \text{Stab}_G(x)$ then $k \cdot x = g \cdot x$. Thus a coset of $\text{Stab}_G(x)$ consists of exactly those group elements that map x to a given element of the orbit $\text{Orb}_G(x)$.

By Proposition 2.29 we know that $|g \text{Stab}_G(x)| = |\text{Stab}_G(x)|$. So for any element $y \in \text{Orb}_G(x)$ there are exactly $|\text{Stab}_G(x)|$ group elements that map x to y . By Corollary 2.28, these left cosets completely partition G , and so there must be exactly $|G|/|\text{Stab}_G(x)|$ distinct elements in the orbit $\text{Orb}_G(x)$. Therefore $|G| = |\text{Orb}_G(x)| \times |\text{Stab}_G(x)|$ for any $x \in X$, as claimed. $\square$

Here's another simple example.

Example 6.26 Let $\mathbb{Z}_7$ act on the circle $C = \{e^{i\theta} : 0 \leq \theta < 2\pi\}$ by $n \cdot e^{i\theta} = e^{i(\theta + 2\pi/7)}$. Then the stabiliser $\text{Stab}_{\mathbb{Z}_7}(x)$ of some point $x = e^{i\phi} \in C$ is just the trivial subgroup $\{0\}$, so $|\text{Stab}_{\mathbb{Z}_7}(x)| = 1$. Since $|\mathbb{Z}_7| = 7$, by Orbit–Stabiliser Theorem, the orbit $\text{Orb}_{\mathbb{Z}_7}(x)$ should consist of seven points.

And this is exactly what happens. The orbit $\text{Orb}_{\mathbb{Z}_7}(x) = \text{Orb}_{\mathbb{Z}_7}(e^{i\phi})$ is equal to the set $\{e^{i\phi + 2k\pi/7} : 0 \leq k < 7\}$, which consists of the point $x = e^{i\phi}$ together with six other points spaced at intervals of $\frac{2\pi}{7}$ around the unit circle C .

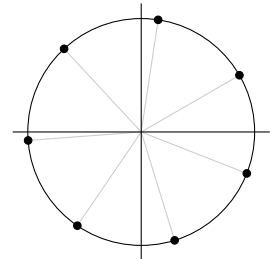


Figure 6.6: An orbit of the $\mathbb{Z}_7$ -action on $C = \{e^{i\theta} : 0 \leq \theta < 2\pi\}$

If we apply the Orbit–Stabiliser Theorem to the conjugation action, then something interesting and useful happens. The orbit $\text{Orb}_G(g)$ of some element $g \in G$ is simply the conjugacy class of g . The stabiliser $\text{Stab}_G(g)$ of g , meanwhile, is composed of those elements h of G that fix g under conjugation; that is, $\text{Stab}_G(g) = \{h \in G : h * g * h^{-1} = g\}$. We’ve met this before, however: it’s the centraliser $Z_G(g)$ of the element g . So, applying the Orbit–Stabiliser Theorem to the conjugation action we get the following corollary:

Corollary 6.27 *Let g be an element of some group G , and let $[g]$ denote the conjugacy class of g . Then $|[g]| = |G : Z_G(g)|$.*

If G is abelian then obviously any group element g commutes with every other element, and so $Z_G(g) = G$. Therefore $|G : Z_G(g)| = 1$, which tells us that the conjugacy class of g consists just of g itself, something we already know from Proposition 3.8. More generally, suppose that G isn’t necessarily abelian, but that g lies in the centre $Z(G)$ and thus commutes with every element of G . Then $Z_G(g) = G$ and so again the conjugacy class of g consists only of g .

Recall from Proposition 3.10 that a subgroup $H \leq G$ is normal if and only if it is a union of conjugacy classes. We can use this to construct a different proof that the alternating group A_5 is simple:

Proposition 6.28 *The alternating group A_5 is simple; that is, it has no proper, nontrivial normal subgroups.*

Proof Recall that A_5 consists of all even permutations of the set $X_5 = \{1, 2, 3, 4, 5\}$. With a little thought and experimentation, we can see that any such permutation must be either the identity ι , a pair of disjoint transpositions of the form $(x_1 x_2)(x_3 x_4)$, a three-cycle of the form $(x_1 x_2 x_3)$ or a five-cycle of the form $(x_1 x_2 x_3 x_4 x_5)$ where x_1, x_2, x_3, x_4, x_5 are distinct elements of X_5 .

The identity permutation ι forms a conjugacy class

$$c_1 = \{\iota\}$$

on its own.

Consider two permutations $\sigma = (x_1 x_2)(x_3 x_4)$ and $\tau = (y_1 y_2)(y_3 y_4)$ in A_5 . Let $\alpha, \beta: X_5 \rightarrow X_5$ be bijections (and hence permutations in S_5) such that

$$\begin{aligned} \alpha: x_1 \mapsto y_1, \quad x_2 \mapsto y_2, \quad x_3 \mapsto y_3, \quad x_4 \mapsto y_4, \quad x_5 \mapsto y_5; \\ \beta: x_1 \mapsto y_2, \quad x_2 \mapsto y_1, \quad x_3 \mapsto y_3, \quad x_4 \mapsto y_4, \quad x_5 \mapsto y_5. \end{aligned}$$

Notice that $\alpha\sigma\alpha^{-1} = \tau$, and also that $\beta\sigma\beta^{-1} = \tau$. Since $\beta = \alpha(x_1 x_2)$, either α or β must have even parity, with the other one being an odd permutation. Therefore there exists an even permutation in A_5 which

conjugates σ to τ . Hence the permutations of the form $(x_1 x_2)(x_3 x_4)$ form a single conjugacy class

$$c_2 = \{(12)(34), (12)(35), (12)(45), (13)(24), (13)(25), \\ (13)(45), (14)(23), (14)(25), (14)(35), (15)(23), \\ (15)(24), (15)(34), (23)(45), (24)(35), (25)(34)\}$$

with fifteen elements.

Similarly, consider two three-cycles $\sigma = (x_1 x_2 x_3)$ and $\tau = (y_1 y_2 y_3)$. Let $\alpha, \beta \in S_5$ be defined by

$$\begin{aligned} \alpha: x_1 \mapsto y_1, \quad x_2 \mapsto y_2, \quad x_3 \mapsto y_3, \quad x_4 \mapsto y_4, \quad x_5 \mapsto y_5; \\ \beta: x_1 \mapsto y_1, \quad x_2 \mapsto y_2, \quad x_3 \mapsto y_3, \quad x_4 \mapsto y_5, \quad x_5 \mapsto y_4. \end{aligned}$$

Observe again that $\alpha\sigma\alpha^{-1} = \beta\sigma\beta^{-1} = \tau$, and that since $\beta = \alpha(x_4 x_5)$ either α or β must be an even permutation and hence in A_5 . So again we see that any three-cycle in A_5 must be conjugate to any other one, and therefore the three-cycles form a conjugacy class

$$c_3 = \{(123), (124), (125), (132), (134), \\ (135), (142), (143), (145), (152), \\ (153), (154), (234), (235), (243), \\ (245), (253), (254), (345), (354)\}$$

with twenty elements.

The remaining 24 elements of A_5 are cycles of length 5, but it turns out that these split neatly into two conjugacy classes, each with twelve elements.

$$\begin{aligned} c_4 = \{(12345), (12453), (12534), (13254), \\ (13425), (13542), (14235), (14352), \\ (14523), (15243), (15324), (15432)\} \\ c_5 = \{(12354), (12435), (12543), (13245), \\ (13452), (13524), (14253), (14325), \\ (14532), (15234), (15342), (15423)\} \end{aligned}$$

To see that c_4 and c_5 are disjoint, suppose that $\sigma = (12345)$ and $\tau = (12354)$, and let $\alpha = (45)$. Then $\alpha\sigma\alpha^{-1} = \tau$. But α has odd parity and is therefore not in A_5 . Suppose instead that there exists $\beta \in A_5$ such that $\beta\sigma\beta^{-1} = \tau$. Then $\alpha\sigma\alpha^{-1} = \beta\sigma\beta^{-1}$, which means that $\alpha^{-1}\beta\sigma = \sigma\alpha^{-1}\beta$. Hence $\alpha^{-1}\beta$ commutes with σ . If $\beta \in A_5$ then it must have even parity, so $\alpha^{-1}\beta$ must have odd parity. So such an even permutation β can exist if and only if there is an odd permutation $\gamma \in S_5$ that commutes with σ . This means that

$$\gamma\sigma\gamma^{-1} = \gamma(12345)\gamma^{-1} = (\gamma(1)\gamma(2)\gamma(3)\gamma(4)\gamma(5)) = (12345).$$

So if $\gamma(1) = i$ then $\gamma(2) = i+1 \pmod{5}$, $\gamma(3) = i+2 \pmod{5}$ and so on. Therefore γ must be a power of σ , and hence an even permutation. Thus σ and τ can't be conjugate.

It only remains to show that c_4 and c_5 are conjugacy classes of 5-cycles in A_5 rather than unions of smaller classes. There are 24 different 5-cycles of the form $(1\ x_2\ x_3\ x_4\ x_5)$, determined by the 24 cycles $(x_2\ x_3\ x_4\ x_5)$ of length 4. Half of these 4-cycles are even, and half of them are odd. The even ones determine the 12 conjugate 5-cycles in c_4 , and the odd ones determine the 12 conjugate 5-cycles in c_5 .

So, we now have the five conjugacy classes in A_5 , and we find that

$$|c_1| = 1, \quad |c_2| = 15, \quad |c_3| = 20, \quad |c_4| = 12, \quad |c_5| = 12.$$

Now suppose that N is a normal subgroup of A_5 . Then by Lagrange's Theorem⁹ $|N|$ must be a factor of $|A_5| = 60$, and by Proposition 3.10 we know that N must be a union of some or all of the conjugacy classes c_1 to c_5 . Also, N must contain the identity permutation ι and hence the class c_1 . But the only factors of 60 that can be formed as a sum of the number 1 together with some or all of the numbers 15, 20, 12 and 12 are 1 and 60. So N must be either the trivial subgroup or A_5 itself, and therefore A_5 is simple. $\square$

⁹ Theorem 2.30, page 54.

It is pointless to do with more that which can be done with fewer.

— William of Ockham (c.1287–1347),
Summa Logicae (c.1323)

6.3 Counting

THE MACHINERY OF GROUP ACTIONS, ORBITS AND STABILISERS has a number of useful applications in combinatorics, and in this section we will study two particular powerful results: Burnside's Lemma and Pólya's Enumeration Theorem. The former is named after the British group theorist William Burnside (1852–1927), although he attributed it to the German mathematician Ferdinand Georg Frobenius (1849–1917), and the French mathematician Augustin-Louis Cauchy (1789–1857) had discovered it even earlier. Given that Burnside didn't originally discover the Lemma, and that it's not the only lemma or theorem named after him, it's been suggested that a more accurate description is "Not Burnside's Lemma".¹⁰ Meanwhile, Pólya's Enumeration Theorem was originally discovered in 1927 by the American mathematician John Redfield (1879–1944) and rediscovered independently ten years later by the Hungarian mathematician George Pólya (1887–1985), after whom it was subsequently named.¹¹

Both Burnside's Lemma and Pólya's Enumeration Theorem are useful for solving enumeration problems related to counting distinct colourings (or analogous configurations) of objects modulo some notion of

¹⁰ PM Neumann, *A lemma that is not Burnside's*, The Mathematical Scientist 4.2 (1979) 133–141

¹¹ This is a common phenomenon in science, known variously as **Stigler's Law of Eponymy** ("No scientific discovery is named after its original discoverer") or **Boyer's Law** ("Mathematical formulas and theorems are usually not named after their discoverers").¹² Naturally, neither of these laws are named after the people who actually coined them either.

¹² HC Kennedy, *Who discovered Boyer's Law?*, American Mathematical Monthly 79.1 (1972) 66–67

symmetry. More generally, both are concerned with counting orbits of a given group action.

For example, suppose that we have a square picture frame and pots of black and white paint. How many different ways are there of painting the four edges of this frame either black or white?

At first sight we might be tempted to say 16, because each of the four edges can be painted one of two colours, so the number of possible configurations is $2^4 = 16$. These are depicted in Figure 6.7.

However, some of these are really the same as each other, just rotated clockwise or anticlockwise by a quarter- or half-turn. If we collect together those coloured frames that only differ by a rotation, we find there are really just six different ways of painting the edges of a square frame:

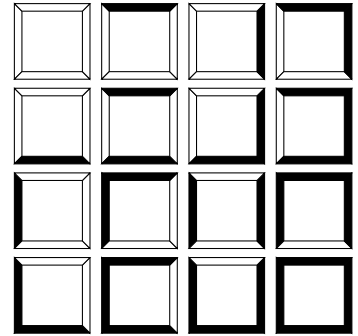
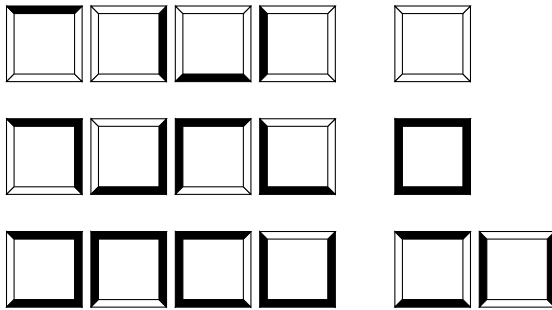


Figure 6.7: Coloured frames



What we have here is a set X consisting of 16 coloured picture frames, together with an action of the rotation group $R_4 \cong \mathbb{Z}_4$. In other words, given any element of X , we have a well-defined way of making another (possibly identical) element of X by rotating it through an angle of $0, \frac{\pi}{2}, \pi$ or $\frac{3\pi}{2}$. The six subsets shown above are exactly the orbits of this action.

So, our question now reduces to one about counting the orbits of a group action, and the answer is given by the following result.

Theorem 6.29 (Burnside's Lemma) *Let G be a finite group acting on a set X . Then the number of distinct orbits of this action is given by*

$$\frac{1}{|G|} \sum_{g \in G} |\text{Fix}_g(X)|,$$

where $\text{Fix}_g(X) = \{x \in X : g \cdot x = x\}$ is the subset of elements of X which are invariant under the action of the element $g \in G$.

Proof First we count the number of pairs $(g, x) \in G \times X$ for which $g \cdot x = x$. There are two obvious ways of doing this. One way is to sum the number of elements in each set $\text{Fix}_g(X)$ for each $g \in G$; this gives the sum $\sum_{g \in G} |\text{Fix}_g(X)|$. The other way is to sum the elements



Wikimedia Commons

William Burnside (1852–1927) was born in London and educated at Christ's Hospital after his father's death in 1858. At school he excelled in mathematics and won a scholarship to St John's College, Cambridge in 1871, but transferred to Pembroke College in 1873, to improve his chances of a place in the college rowing team. He graduated in 1875 with the second highest marks in his year, winning the title of Second Wrangler and a prestigious Smith's Prize.

He remained in Cambridge for another ten years, as a lecturer and Fellow of Pembroke, but in 1885 accepted the post of Professor of Mathematics at the Royal Naval College in Greenwich, where he stayed for the rest of his career, despite being offered the post of Master of Pembroke in 1903.

After early work on elliptic functions, he became interested in hydrodynamics, which led him to study complex function theory and finite group theory. His work on the latter resulted in the publication of his influential book *The Theory of Groups of Finite Order* in 1897, the first treatise on group theory in English. His many contributions to the field include Burnside's $p^a q^b$ Theorem and the Burnside Problem. Later in his career he became interested in probability, writing several papers on the subject, as well as a book entitled *Theory of Probability*, which was published posthumously in 1928.

He was elected a Fellow of the Royal Society in 1893 and awarded the De Morgan Medal of the London Mathematical Society in 1899, serving on the latter's council until 1917 and as its president between 1906 and 1908.

in the stabiliser $\text{Stab}_G(x)$ as x ranges over all the elements in X ; this gives the sum $\sum_{x \in X} |\text{Stab}_G(x)|$. These two sums must be equal, hence

$$\sum_{g \in G} |\text{Fix}_g(X)| = \sum_{x \in X} |\text{Stab}_G(x)|. \quad (6.9)$$

If we let $X_1, \dots, X_k$ be the distinct orbits of the G -action on X , this second sum can be rewritten as $\sum_{i=1}^k \sum_{x \in X_i} |\text{Stab}_G(x)|$.

If two elements x_1 and x_2 lie in the same orbit, then $\text{Orb}_G(x_1) = \text{Orb}_G(x_2)$ and $\text{Stab}_G(x_1) = \text{Stab}_G(x_2)$. So, we can choose a specific representative element $x_i \in X_i$ for each orbit X_i . Then

$$\sum_{x \in X_i} |\text{Stab}_G(x)| = |X_i| \times |\text{Stab}_G(x_i)| = |\text{Orb}_G(x_i)| \times |\text{Stab}_G(x_i)|.$$

But by the Orbit–Stabiliser Theorem (Theorem 6.25) we know that this is equal to $|G|$. Substituting this into (6.9) we get

$$\sum_{g \in G} |\text{Fix}_g(X)| = \sum_{x \in X} |\text{Stab}_G(x)| = k|G|,$$

where k is the number of orbits. Hence

$$k = \frac{1}{|G|} \sum_{g \in G} |\text{Fix}_g(X)|$$

as required. $\square$

Burnside's Lemma says that the number of orbits of a G -action is the mean number of elements of X left fixed by elements of G .

Applying this result to our picture frame problem, we need to count the number of frames which are unchanged by a given element of $R_4 = \{e, r, r^2, r^3\} \cong \mathbb{Z}_4$. The identity e leaves everything fixed, so $|\text{Fix}_e(X)| = 16$. The anticlockwise rotation r and the clockwise rotation r^3 both leave the following frames unchanged:



The half-turn r^2 leaves four frames unchanged:



So, applying Burnside's Lemma we find that this action has

$$\frac{1}{4}(16 + 2 + 2 + 4) = 6$$

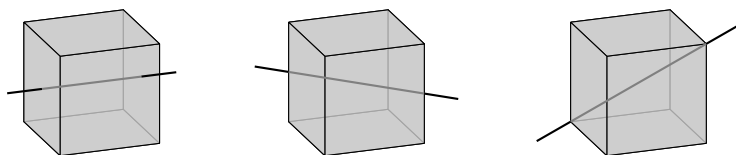
orbits. This agrees with the answer we found earlier by inspection.

Here's another example:

Example 6.30 An ordinary six-sided die is a cube with each face numbered from 1 to 6. We'll consider three variants on this theme and work out how many configurations there are in each case.

Firstly, consider the general case where we number each face independently, allowing more than one face to have the same number. There are $6^6 = 46656$ possible choices, but some of these will be just rotated versions of each other.

The direct isometry group $\text{Isom}^+(\square^3)$ consists of all the orientation-preserving isometries of the cube $\square^3$. There are 24 of these: the identity e , six order-2 rotations t about axes passing through the midpoints of opposite edges, three order-2 rotations r^2 about axes passing through the centres of opposite faces, eight order-3 rotations s about axes passing through opposite corners, and six order-4 rotations r about axes passing through the centres of opposite faces.



The identity e obviously fixes all 46656 possible cubes, the six type- t rotations fix $6^3 = 216$ cubes, the three type- r^2 rotations fix $6^4 = 1296$ cubes, the eight type- s rotations fix $6^2 = 36$ cubes, and the six type- r rotations fix $6^3 = 216$ cubes.

So, Burnside's Lemma tells us that there are

$$\frac{1}{24}(46656 + 6 \times 216 + 3 \times 1296 + 8 \times 36 + 6 \times 216) = \frac{53424}{24} = 2226$$

distinct ways of numbering the faces of the die.

But a conventional die uses each number only once, so how many distinct dice can we construct with this property? Well, instead of $6^6 = 46656$ possible configurations we have only $6! = 720$: we have six possible choices for the top face, five remaining for the front face, four for the left face and so on. The identity fixes all 720 of these, but none of the other 23 direct isometries fix any. So Burnside's Lemma tells us that there are $\frac{720}{24} = 30$ distinct dice of this type.

Actually, most conventional six-sided dice also satisfy the condition that opposite faces add up to seven: the 6 face is opposite the 1 face, the 5 opposite the 2 and the 4 opposite the 3. How many different dice are there with this property too?

We have six choices of number for the top face, and when we've made that choice we know what the number on the bottom face has to be. Next we choose one of the four remaining numbers for the front face,



Oberwolfach Photo Collection / Mathematische Gesellschaft (Hamburg)

The German mathematician **Ferdinand Georg Frobenius** (1849–1917)

at which point we have no choice about what number goes on the back face. Finally, we have two remaining numbers to choose from for the left face, which leaves only one for the right face. Therefore there are $6 \times 4 \times 2 = 48$ possible configurations satisfying these criteria. As before, the identity fixes all 48 and no other isometry fixes any of them. By Burnside's Lemma there are exactly $\frac{48}{24} = 2$ such dice.

Burnside's Lemma enables us to count orbits of a group action, and as we've seen this can help us count colourings or other configurations of some geometric object modulo some equivalence relation such as rotational symmetry. But in general the task of counting the number of elements in $\text{Fix}_g(X)$ can become cumbersome and time-consuming. For something relatively simple like a square picture frame or a cubic die, the process is fairly straightforward, but it becomes less so for more complicated objects. For example, a regular dodecahedron has thirty edges and a direct isometry group with sixty elements, so counting its distinct edge-colourings is going to be a lengthier undertaking than the examples we've looked at so far.

In the picture frame example it was fairly straightforward to work out the orders of $\text{Fix}_g(X)$ for two colours:

$$\begin{aligned} |\text{Fix}_e(X)| &= 2^4 = 16, & |\text{Fix}_r(X)| &= 2^1 = 2, \\ |\text{Fix}_{r^2}(X)| &= 2^2 = 4, & |\text{Fix}_{r^3}(X)| &= 2^1 = 2. \end{aligned}$$

What if we have three colours at our disposal: black, white and grey?

There are $3^4 = 81$ possible colourings, of which all are fixed by the identity e , 3 are fixed by r and r^2 , and 9 are fixed by r^2 , giving the following values for $|\text{Fix}_g(X)|$:

$$\begin{aligned} |\text{Fix}_e(X)| &= 3^4 = 81, & |\text{Fix}_r(X)| &= 3^1 = 3, \\ |\text{Fix}_{r^2}(X)| &= 3^2 = 9, & |\text{Fix}_{r^3}(X)| &= 3^1 = 3. \end{aligned}$$

There is an obvious pattern here, and with a little thought we can see that for k colours the answer is going to be

$$\begin{aligned} |\text{Fix}_e(X)| &= k^4, & |\text{Fix}_r(X)| &= k^1 = k, \\ |\text{Fix}_{r^2}(X)| &= k^2, & |\text{Fix}_{r^3}(X)| &= k^1 = k. \end{aligned}$$

Therefore there are $\frac{1}{4}(k^4 + k^2 + 2k)$ rotationally distinct ways of colouring our square picture frame with k different colours.

The key is to study how the isometry group acts on the object we're colouring. In this case, the rotation group $\text{Isom}^+(\square^2) = R_4 \cong \mathbb{Z}_4$ permutes the four sides of the square: this yields a homomorphism $\phi: R_4 \rightarrow S_4$ which we can write down in cycle form as follows:

$$\begin{aligned} e &\longmapsto \iota = (1)(2)(3)(4), & r &\longmapsto (1\,2\,3\,4), \\ r^2 &\longmapsto (1\,3)(2\,4), & r^3 &\longmapsto (1\,4\,3\,2). \end{aligned}$$

Here we've numbered the edges of the frame anticlockwise with the numbers 1–4. For what follows it's helpful to also include cycles of length 1, which for conciseness we don't usually bother doing.

To each cycle we associate a symbol, its **type**, as follows.

Definition 6.31 Suppose that a permutation $\sigma \in S_n$ can be decomposed as a product of disjoint cycles, of which a_1 have length 1, a_2 have length 2, and so on, where $a_1, a_2, \dots$ are non-negative integers.

Then the **type** of the permutation σ is the partition $[1^{a_1}, 2^{a_2}, \dots, n^{a_n}]$ of n , where $\sum_{i=1}^n a_i i = a_1 + 2a_2 + \dots + na_n = n$.

The **cycle symbol** cs_σ of σ is the formal monomial $x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$.

So, the cycle symbols of the elements of R_4 , regarded as permutations in S_4 via the action homomorphism ϕ , are

$$\text{cs}_e = x_1^4, \quad \text{cs}_r = x_4, \quad \text{cs}_{r^2} = x_2^2, \quad \text{cs}_{r^3} = x_4.$$

The cycle symbol $\text{cs}_g = \text{cs}_g(x_1, \dots, x_n)$ is a monomial in the formal symbols $x_1, \dots, x_n$. To count the k -colourings fixed by a group element g , we evaluate $\text{cs}_g(x_1, \dots, x_n)$ setting $x_1 = \dots = x_n = k$. We can see why this works by looking at what happens in a particular case.

The group element r^2 has cycle symbol $\text{cs}_{r^2} = x_2^2$, because it corresponds to the permutation (13)(24). If sides 1 and 3 are the same colour as each other, and sides 2 and 4 are also the same colour as each other, then this permutation will leave the frame unchanged. So the disjoint cycles group the edges together in identically-coloured cohorts, each of which can be coloured any of the k available colours. So we have k choices of colour for each of these groupings; that is, k choices for each disjoint cycle in the permutation. More generally, for a permutation that decomposes into m disjoint cycles, there are k^m colourings of our chosen object that are invariant under that permutation.

Definition 6.32 Let a finite group G act on a finite set X via an action homomorphism $\phi: G \rightarrow \text{Sym}(X) \cong S_n$. Then the **cycle index** of the action is the polynomial

$$C_G(x_1, \dots, x_n) = \frac{1}{|G|} \sum_{g \in G} \text{cs}_g(x_1, \dots, x_n).$$

Different books adopt different notation for the cycle index, such as $Z(G)$, P_G and ζ_G . Here we've settled on C_G , but you should be aware that there is no strong consensus. Also, some books use $C_G(S)$ to denote the centraliser of S in G (Definition 4.51) and $C(G)$ to denote the centre of G (Definition 3.16).

By the argument above, we get the following corollary to Burnside's Lemma:

Corollary 6.33 (Cycle Index Theorem) *Let X be an object with n components, each of which is assigned one of k colours. The number of distinct k -colourings of X modulo an action of G is given by*

$$C_G(k, \dots, k) = \frac{1}{|G|} \sum_{g \in G} \text{cs}_g(k, \dots, k).$$

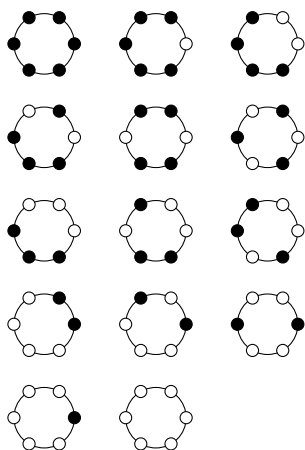


Figure 6.8: The 14 rotationally distinct 2-coloured 6-bead necklaces

Let's apply this to another colouring problem.

Example 6.34 Consider a circular **necklace** with n coloured but otherwise identical and equally-spaced beads. Up to rotation, how many different necklaces can we make if we have k different colours to choose from?

The general case is a bit involved so first we'll consider the case $n = 6$. We want to count the orbits of the $\mathbb{Z}_6$ -action. The permutation representations and cycle symbols of the elements of $\mathbb{Z}_6$ are

$$\begin{aligned}
 0 &\mapsto \iota = (1)(2)(3)(4)(5)(6), & \text{cs}_0 &= x_1^6, \\
 1 &\mapsto (1\,2\,3\,4\,5\,6), & \text{cs}_1 &= x_6, \\
 2 &\mapsto (1\,3\,5)(2\,4\,6), & \text{cs}_2 &= x_3^2, \\
 3 &\mapsto (1\,4)(2\,5)(3\,6), & \text{cs}_3 &= x_2^3, \\
 4 &\mapsto (1\,5\,3)(2\,6\,4), & \text{cs}_4 &= x_2^3, \\
 5 &\mapsto (1\,6\,5\,4\,3\,2), & \text{cs}_5 &= x_6.
 \end{aligned}$$

The cycle index is therefore

$$C_{\mathbb{Z}_6}(x_1, x_2, x_3, x_4, x_5, x_6) = \frac{1}{6}(x_1^6 + x_6 + x_3^2 + x_2^3 + x_2^3 + x_6)$$

and hence the number of necklaces with two colours of beads is

$$C_{\mathbb{Z}_6}(2, 2, 2, 2, 2, 2) = \frac{1}{6}(2^6 + 2^3 + 2 \times 2^2 + 2 \times 2) = \frac{84}{6} = 14.$$

These are depicted in Figure 6.8.

The general case requires the cycle decomposition of the appropriate permutation representation of $\mathbb{Z}_n$.

Proposition 6.35 *There are*

$$\frac{1}{n} \sum_{d|n} \phi(d) k^{n/d}$$

rotationally distinct k -coloured n -bead necklaces.

Here, ϕ denotes Euler's **totient function**.¹³

Proof Following the method in Example 6.34, we need to work out the cycle symbols cs_i for each $i \in \mathbb{Z}_n$. More precisely, we want the cycle symbols for the images σ_i of each i in the symmetric group S_n .

The permutation σ_i has a cycle of the form $(1\,(1+i)\,(1+2i)\,\dots)$. If i is coprime to n then this cycle will be of length n , and the cycle symbol for σ_i will be x_n . If $h = \gcd(i, n) \neq 1$ then this cycle will be of length n/h and the permutation σ_i will decompose as

$$\begin{aligned}
 \sigma_i &= (1\,(1+i)\,(1+2i)\,\dots)(2\,(2+i)\,(2+2i)\,\dots)\dots \\
 &\quad ((h-1)\,(h-1+i)\,(h-1+2i)\,\dots)
 \end{aligned}$$

and so the cycle symbol of σ_i will be $x_{n/h}^h$.

¹³ Definition 2.44, page 63.

Observe that the length of each cycle must be a factor of n , so the subscripts and superscripts of the cycle symbols for the permutations σ_i must be factors of n . Furthermore, all factors of n will occur in this way.

So the full cycle index for this action will be

$$\frac{1}{n} \sum_{d|n} m_d x_d^{n/d}$$

where m_d is the number of permutations σ_i that decompose as a product of n/d disjoint cycles of length d . To calculate this, we need to know how many integers m satisfy $n/\gcd(m, n) = d$ for a given divisor d .

Such an $m = \gcd(m, n) \times t$ for some t with $\gcd(t, d) = 1$. There are $\phi(d)$ choices of such a t , so $m_d = \phi(d)$ and hence the cycle index is

$$\frac{1}{n} \sum_{d|n} \phi(d) x_d^{n/d}.$$

Setting all the $x_i^j = k$ and applying Corollary 6.33 we get

$$\frac{1}{n} \sum_{d|n} \phi(d) k^{n/d}$$

as claimed. $\square$

We now have some useful machinery for counting colourings or other configurations up to some form of symmetry equivalence. But the results we've seen so far don't really tell us anything about what these colourings actually look like.

Suppose we have a k -coloured n -bead necklace, and let $C = \{c_1, \dots, c_k\}$ be a set of labels representing the k colours.

Then we represent a given colouring of the necklace with a degree- n monomial $c_1^{m_1} c_2^{m_2} \dots c_k^{m_k}$ where $m_1 + \dots + m_k = n$ are non-negative integers. This denotes a colouring where m_1 beads are coloured with colour c_1 , m_2 beads with colour c_2 and so on.

There is often more than one colouring with these parameters, and we denote this by a non-negative integer coefficient in front of the monomial in question.

For example, in the case of the 6-bead necklace in Example 6.34, we had two colours $c_1 = \text{white}$ and $c_2 = \text{black}$. Of the $2^6 = 64$ possible necklaces (ignoring rotation for the moment), one has six white beads, six have five white beads, fifteen have four white beads, twenty have three white beads, fifteen have two white beads, six have a single white bead and one has only black beads. We can represent this situation with the polynomial

$$c_1^6 + 6c_1^5c_2 + 15c_1^4c_2^2 + 20c_1^3c_2^3 + 15c_1^2c_2^4 + 6c_1c_2^5 + c_2^6.$$

Or, using more intuitive names for the colours,

$$w^6 + 6w^5b + 15w^4b^2 + 20w^3b^3 + 15w^2b^4 + 6wb^5 + b^6.$$

This is an example of a **generating function**, a device which has many applications in combinatorics. In this case, it happens to be equal to $(w + b)^6$.

But this isn't quite what we want, because we're really interested in something which encodes the number of rotationally distinct necklaces with a given number of beads of each colour. With that in mind, we modify our generating function accordingly, to get

$$w^6 + w^5b + 3w^4b^2 + 4w^3b^3 + 3w^2b^4 + wb^5 + b^6.$$

From this we can easily read off the details of all the individual rotationally-distinct colourings of the necklace. What we need now is a general method for constructing this generating function from the cycle symbols of the group action.

To see how to do this, we consider each group element individually and look at the monomials corresponding to the necklaces which are invariant under that element:

g	monomials	cs_g
0	$w^6 + 6w^5b + 15w^4b^2 + 20w^3b^3 + 15w^2b^4 + 6wb^5 + b^6$	$= (w+b)^6 \quad x_1^6$
1	$w^6 + b^6$	$= (w^6 + b^6) \quad x_6$
2	$w^6 + 2w^3b^3 + b^6$	$= (w^3 + b^3)^2 \quad x_3^2$
3	$w^6 + 3w^4b^2 + 3w^2b^4 + b^6$	$= (w^2 + b^2)^3 \quad x_2^3$
4	$w^6 + 2w^3b^3 + b^6$	$= (w^3 + b^3)^2 \quad x_3^2$
5	$w^6 + b^6$	$= (w^6 + b^6) \quad x_6$

Adding all of these together and dividing by $|\mathbb{Z}_6| = 6$ we get

$$\frac{1}{6}(6w^6 + 6w^5b + 18w^4b^2 + 24w^3b^3 + 18w^2b^4 + 6wb^5 + b^6)$$

which is exactly the generating function

$$w^6 + w^5b + 3w^4b^2 + 4w^3b^3 + 3w^2b^4 + wb^5 + b^6$$

that we want. If we look carefully at the table above then a pattern suggests itself: a group element with cycle symbol x_i^j contributed a term $(w^i + b^i)^j$ to the generating function. Summing over all group elements and dividing by the order of the group gives us exactly the right generating function. What we've done here is substitute $x_i = (w^i + b^i)$ in the expression for the cycle index $\text{C}_{\mathbb{Z}_6}(x_1, \dots, x_6)$.

This generating function tells us, amongst other things, that there are four different rotationally distinct 2-coloured 6-bead necklaces, the relevant term being $4w^3b^3$.

It turns out that this procedure works in general:

Theorem 6.36 (Pólya's Enumeration Theorem) *Let X be an object with n components, each of which is assigned one of k colours $c_1, \dots, c_k$. The generating function describing the distinct k -colourings of X , modulo an action of some finite group G , is given by*

$$C_G(c_1 + \dots + c_k, c_1^2 + \dots + c_k^2, \dots, c_1^n + \dots + c_k^n) = \frac{1}{|G|} \sum_{g \in G} \text{cs}_g(c_1 + \dots + c_k, c_1^2 + \dots + c_k^2, \dots, c_1^n + \dots + c_k^n).$$

Proof Let $g \in G$ and suppose that the cycle decomposition of the permutation representing g in S_n consists of a single cycle $(a_1 \dots a_n)$. Then $\text{Fix}_g(X)$ consists only of colourings in which each component of X has the same colour. The generating function in this case is

$$c_1^n + \dots + c_k^n$$

and the cycle symbol $\text{cs}_g = x_n$.

Now suppose that the permutation representative for g decomposes as more than one disjoint cycle. We can thus regard X as consisting of several independently-coloured subobjects with respect to this group element g , each corresponding to one of the cycles in the permutation representing g .

A cycle of length m has a generating function of the form

$$c_1^m + \dots + c_k^m$$

by the argument above, and contributes a factor x_m to the cycle symbol cs_g . The cycle symbol cs_g is the product of all of these factors, and the generating function for g is equal to the product of the corresponding generating functions.

Therefore the generating function for the element g , which describes the k -colourings of X which are invariant under the action of g , can be obtained by substituting $x_i = c_1^i + \dots + c_k^i$ in the cycle symbol $\text{cs}_g(x_1, \dots, x_n)$.

In order to find the generating function that takes account of the whole G -action on X , we need to do a little more work.

For a given colour distribution $c_1^{m_1} \dots c_k^{m_k}$, we obtain the number of equivalent colourings by summing the appropriate term from the generating function for each $g \in G$ and then divide by $|G|$.

To get the generating function for all possible colourings, then, we have to take the generating function for all $g \in G$, sum them and divide by $|G|$. This gives the expression

$$\frac{1}{|G|} \sum_{g \in G} \text{cs}_g(c_1 + \dots + c_k, c_1^2 + \dots + c_k^2, \dots, c_1^n + \dots + c_k^n)$$

as claimed. $\square$



ETH Bibliothek, Zurich

George (György) Pólya (1887–1985) was born in Hungary and initially studied language and literature for two years at the University of Budapest, before switching to mathematics and physics. He studied abroad in Vienna and Göttingen, completed his doctorate at Budapest with very little supervision, and then visited Paris in 1914. There, he met the German mathematician Adolf Hurwitz (1860–1919) who arranged an academic post for him in Zürich.

He stayed in Zürich for a number of years, publishing prolifically on a wide range of mathematical topics. With his fellow Hungarian mathematician Gábor Szegő (1895–1985) he wrote an influential two-volume analysis problem book: *Aufgaben und Lehrsätze aus der Analysis*, published in 1925.

In 1940 he moved to the USA, and after two years at Brown University was appointed to a post at Stanford, where he remained for the rest of his life. Retiring in 1953, he was appointed Professor Emeritus, and continued teaching and researching into his nineties.

His work spanned a wide range of topics: analysis, algebra, combinatorics, geometry, probability and geometry in particular. His classic book on mathematical problem solving, *How to Solve It*, first published in 1945, has sold over a million copies and been translated into seventeen languages, and remains recommended reading for any student of mathematics.

Summary

¹⁴ Definition 1.43, page 21.

¹⁵ Example 1.34, page 17.

¹⁶ Definition 1.51, page 24.

¹⁷ Definition 6.1, page 206.

Motivated by several groups discussed in Chapter 1, particularly the dihedral groups D_n ,¹⁴ the general linear groups $GL_n(\mathbb{R})$,¹⁵ and the symmetric groups S_n ,¹⁶ we formulated the concept of an **action** of a group G on a set X :¹⁷ a function $\alpha: G \times X \rightarrow X$ satisfying certain simple criteria. We usually denote the image of a given element $x \in X$ under the action of a specific element $g \in G$ by $g \cdot x$ rather than $\alpha(g, x)$. More precisely, this is a **left action** of G on X ; there is a corresponding notion of a **right action** defined by a function $\beta: X \times G \rightarrow X$, in which we denote the images $x \cdot g$. There is a bijective mapping between left and right actions, so in principle we need only develop the theory for left actions, and the analogous results should hold for right actions as well.

¹⁸ Proposition 6.2, page 208.

A group action $\alpha: G \times X \rightarrow X$ determines a unique homomorphism $\phi: G \rightarrow \text{Sym}(X)$, where ϕ maps a given group element $g \in G$ to the permutation f_g of X defined by $f_g(x) = g \cdot x$ for all $x \in X$.¹⁸ If this homomorphism ϕ is injective, we say that the action is **faithful**; equivalently an action is faithful if, for any distinct $g_1, g_2 \in G$ there is at least one $x \in X$ for which $g_1 \cdot x \neq g_2 \cdot x$.¹⁹ Also, an action is faithful exactly when $\ker \phi$ is trivial.²⁰

¹⁹ Definition 6.4, page 209.

²⁰ Proposition 4.35, page 114.

The **trivial action** is defined by $g \cdot x = x$ for all $g \in G$ and $x \in X$; this corresponds to the homomorphism $G \rightarrow \text{Sym}(X)$ where every element of G maps to the identity homomorphism $\iota \in \text{Sym}(X)$.²¹ The trivial action is not faithful except when G is the trivial group $\{e\}$.

²¹ Example 6.6, page 209.

The action of the full symmetric group $\text{Sym}(X)$ on a set X is faithful, and corresponds to the identity homomorphism $\text{id}: \text{Sym}(X) \rightarrow \text{Sym}(X)$.²² For any subgroup $G \leq \text{Sym}(X)$ we can construct a faithful action of G on X by composing the canonical inclusion map $i: G \hookrightarrow \text{Sym}(X)$ with this symmetric group action. In particular, this process yields a faithful action of the alternating group $\text{Alt}(X) < \text{Sym}(X)$ on X , of A_n on the finite set $X_n = \{1, \dots, n\}$,²³ and a faithful action of the special orthogonal group $SO_n(\mathbb{R})$ on $\mathbb{R}^n$.²⁴

²² Example 6.5, page 209.

²³ Example 6.7, page 210.

²⁴ Example 6.8, page 210.

²⁵ Theorem 2.13, page 47.

By Cayley's Theorem²⁵ any group G can be regarded as a subgroup of $\text{Sym}(G)$, the permutation group of its underlying set. This viewpoint yields a faithful action of G on itself: the **left or right regular action**.²⁶

²⁶ Definition 6.9, page 211.

The group conjugation operation introduced in Chapter 3 gives another action of a group on itself: the **conjugation action** in which $g \cdot h = hg = ghg^{-1}$ for any $g, h \in G$.²⁷ This action is not in general faithful. If G is abelian then the conjugation action is trivial.²⁸

²⁷ Definition 6.11, page 211.

²⁸ Proposition 6.13, page 212.

Unfaithful actions have nontrivial kernels, and thereby determine

nontrivial normal subgroups of the given group G . We can take the quotient G by this nontrivial kernel K and obtain a faithful G/K -action on our chosen set.²⁹

²⁹ Example 6.15, page 212.

An action of a group G on a set X determines an equivalence relation on X : for any $x, y \in X$ we say that $x \sim y$ exactly when there is an element $g \in G$ such that $g \cdot x = y$. This relation partitions X into equivalence classes, which we call **orbits**.³⁰ The orbit of an element $x \in X$ is denoted $\text{Orb}_G(x)$.

³⁰ Definition 6.16, page 213.

The orbits of the canonical $SO_3(\mathbb{R})$ -action on $\mathbb{R}^3$ are concentric spheres around the origin, and the origin itself.³¹ The usual action of D_3 on an equilateral triangle has three different types of orbit: the centre of the triangle, orbits comprising triples of points on one of the triangle's axes of symmetry, and orbits consisting of six points off the axes of symmetry.³² The trivial action of a group G on a set X partitions X into single-element subsets of X .³³ The orbits of the conjugation action of a group G on itself are exactly the conjugacy classes of G .³⁴

³¹ Example 6.17, page 214.

³² Example 6.18, page 214.

³³ Example 6.19, page 214.

³⁴ Example 6.20, page 215.

An action with only one orbit is said to be **transitive**.³⁵ The left (or right) regular action of a group G on itself is transitive.³⁶

³⁵ Definition 6.21, page 215.

³⁶ Proposition 6.22, page 215.

The **stabiliser** of an element $x \in X$ under a given G -action is the subset of G consisting of those elements that leave x fixed; that is, those $g \in G$ for which $g \cdot x = x$.³⁷ This subset, denoted $\text{Stab}_G(x)$, is a subgroup of G ,³⁸ and is sometimes called the **isotropy subgroup** of x . If G is a finite group acting on a (not necessarily finite) set X , then $|\text{Orb}_G(x)| \times |\text{Stab}_G(x)| = |G|$ for any $x \in X$; this fact is called the **Orbit–Stabiliser Theorem**.³⁹ In the case of the conjugation action of G on itself, the orbit $\text{Orb}_G(g)$ of some element $g \in G$ is just the conjugacy class $[g]$ of g , and the stabiliser $\text{Stab}_G(g)$ is the centraliser $Z_G(g)$ of g in G .⁴⁰ Hence, by the Orbit–Stabiliser Theorem, we have $|[g]| = |G : Z_G(g)|$; that is, the number of elements of the conjugacy class of g is equal to the index of its centraliser in G .⁴¹ We can use this fact to show that the alternating group A_5 is simple.⁴²

³⁷ Definition 6.23, page 216.

³⁸ Proposition 6.24, page 216.

³⁹ Theorem 6.25, page 217.

⁴⁰ Definition 4.51, page 122.

⁴¹ Corollary 6.27, page 218.

⁴² Proposition 6.28, page 218.

An important result known as **Burnside's Lemma** enables us to count the orbits of a finite group action. If we denote by $\text{Fix}_g(X)$ the subset of X consisting of those elements fixed by an element $g \in G$ (that is, those $x \in X$ for which $g \cdot x = x$) then the number of distinct orbits is given by $\frac{1}{|G|} \sum_{g \in G} |\text{Fix}_g(X)|$.⁴³ We can use this result to solve certain types of problem in enumerative combinatorics: for example, counting the number of six-sided dice satisfying various properties.⁴⁴

⁴³ Theorem 6.29, page 221.

⁴⁴ Example 6.30, page 223.

Suppose that a permutation $\sigma \in S_n$ can be decomposed as a product of disjoint cycles, a_1 with length 1, a_2 with length 2, and so forth, for some non-negative integers $a_1, a_2, \dots$. The **type** of σ is the partition $[1^{a_1}, 2^{a_2}, \dots, n^{a_n}]$, where $a_1 + 2a_2 + \dots + na_n = n$. The **cycle symbol**

⁴⁵ Definition 6.31, page 225.

cs_σ is the formal monomial $x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}$.⁴⁵

⁴⁶ Definition 6.32, page 225.

Given a finite group G acting on a finite set X via an action homomorphism $\phi: G \rightarrow \text{Sym}(X) \cong S_n$, the **cycle index** of the action is the polynomial $C_G(x_1, \dots, x_n) = \frac{1}{|G|} \sum_{g \in G} cs_g(x_1, \dots, x_n)$.⁴⁶ Now consider an object with n components, each of which is assigned one of k colours. Then the **Cycle Index Theorem**, a corollary to Burnside's Lemma, states that the number of distinct k -colourings of X modulo a given G -action is equal to $C_G(k, \dots, k)$.⁴⁷

⁴⁷ Corollary 6.33, page 225.

For example, we can use the Cycle Index Theorem to count the number of different k -colourings of an n -bead necklace.^{48,49}

⁴⁸ Example 6.34, page 226.

⁴⁹ Proposition 6.35, page 226.

⁵⁰ Theorem 6.36, page 229.

We can obtain more detailed information about these colourings by constructing a **generating function**. **Pólya's Enumeration Theorem**⁵⁰ says that the generating function for the k -colouring problem on an n -component object X , modulo an action by a finite group G , is given by the cycle index $C_G(c_1 + \dots + c_k, c_1^2 + \dots + c_k^2, \dots, c_1^n + \dots + c_k^n)$.

References and further reading

- J H Conway, ND Elkies, and J L Martin, *The Mathieu group M_{12} and its pseudogroup extension M_{13}* , Experimental Mathematics 15.2 (2006) 223–236
A detailed discussion of the construction of the Mathieu group M_{12} .
- A S Crans, T M Fiore, and R Satyendra, *Musical actions of dihedral groups*, The American Mathematical Monthly 116.6 (2009) 479–495
An investigation of the applications of group theory to music. The dihedral group D_{12} acts musically by **inversion** and **transposition**, and also via three invertible operations on musical triads: the **parallel** operation P mapping a major triad to the corresponding minor triad, the **leading tone exchange** L which raises the root note by a semitone, and the **relative** operation R mapping a major triad to its relative minor.

Exercises

- 6.1 Show that $C_{G \times H} = C_G C_H$ for any groups G and H .
- 6.2 Suppose that the necklaces in Example 6.34 can also be flipped over. That is, replace the $\mathbb{Z}_6$ -action with an appropriate D_6 -action. Configurations of this type are sometimes called **bracelets**. How many distinct 2-coloured 6-bead bracelets are there?
- 6.3 Modify or extend the proof of Proposition 6.35, replacing $\mathbb{Z}_n$ with D_n , to calculate the number of k -coloured n -bead bracelets.
- 6.4 Convince yourself that A_4 is 2-transitive, and that A_5 is 3-transitive.
- 6.5 Try to construct a Steiner system of type $S(2, 3, 4)$ and convince yourself that no such system exists.

7 Finite groups

IN THIS CHAPTER we will study a number of important results related to the classification of finite groups. Recall that Lagrange's Theorem¹ says that for any finite group G , the order $|H|$ of any subgroup $H \leq G$ must divide $|G|$. But the converse doesn't always hold: for example, $|A_4| = 12$ but A_4 has no subgroup of order 6. Cauchy's Theorem² is a partial converse, ensuring the existence of a (cyclic) subgroup of order p if p is a prime factor of $|G|$.

First we will investigate some important theorems due to the Norwegian mathematician Ludwig Sylow (1832–1918),³ that extend Cauchy's Theorem to p -**subgroups**: subgroups of order p^n .

Next we will study how subgroups can be nested inside each other (like algebraic *matryoshka* dolls), and prove some important results about these **series** of subgroups, such as Schreier's Refinement Theorem⁴ and the Jordan–Hölder Theorem.⁵ Some of this work will lead into our study of Galois Theory in Chapter 11.

After that, we will introduce the **semidirect product** $G = H \rtimes K$ of two groups H and K , a generalisation of the direct product $H \times K$, which will lead us to the study of **group extensions**: a general method of constructing larger groups from smaller ones.

Finally, we will use some of these techniques to classify, up to isomorphism, all the groups with order less than 32.

The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land, to add something to the extent and the solidity of our possessions.

— Thomas Huxley (1825–1895),
On the Reception of the Origin of Species,
in: Francis Darwin, *The Life and Letters of Charles Darwin* (1887) II 204

¹ Theorem 2.30, page 54.

² Theorem 2.37, page 57.

³ Pronounced, approximately, “see-loff” rather than “sigh-low”.

⁴ Theorem 7.27, page 248.

⁵ Theorem 7.34, page 254.

7.1 Sylow's Theorems

WE WANT TO FIND A PARTIAL CONVERSE TO LAGRANGE'S THEOREM; that is, for a finite group G , we want to find which factors of $|G|$ are represented as the order of some subgroup of G . Cauchy's Theorem gives one answer to this question: it guarantees that at least one (cyclic) subgroup of order p exists for any prime factor p of $|G|$. But can we do better than this?

As every educated person knows the Pythagorean Theorem so does every mathematician speak of Abel's Theorem and of Sylow's Theorem.

— Georg Frobenius (1849–1917),
quoted in: G A Miller,
Professor Ludvig Sylow,
Science 49:1256 (1919) 85



Oberwolfach Photo Collection / Wolfgang Gaschütz

Born in Christiania (now Oslo), the eldest of ten siblings, **Peter Ludvig Mejdell Sylow** (1832–1918) studied at the University of Christiania, qualifying as a mathematics and science teacher in 1856. No suitable university post being available, he taught in secondary schools for the next forty-two years.

He wasn't a very good schoolteacher: he found it difficult to keep order, and lacked enthusiasm for the elementary topics he was required to teach.

In 1861 he visited Paris and Berlin, learning of recent advances in mechanics, geometry and algebra. The next year he lectured at the University of Christiania, covering for his colleague Ole Jacob Broch (1818–1889) who had been elected to parliament. Broch was re-elected in 1865, but Sylow's school refused to grant him further leave.

Nevertheless, Sylow continued to pursue mathematical research: in 1872 he published his celebrated theorems on orders of subgroups, and between 1873 and 1881 he and Sophus Lie (1842–1899) published the collected works of their fellow Norwegian mathematician Niels Henrik Abel (1802–1829).

Lie eventually succeeded in getting him appointed to a chair at Christiania in 1898, at the advanced age of 65. He occupied this post for a further twenty years until his death in 1918, successfully undertaking further research in group theory and elliptic functions.

The answer is yes, due to three theorems originally proved by the Norwegian mathematician Ludwig Sylow (1832–1918), and which we will now study.

A good place to start is with our first counterexample A_4 which, as we saw in Example 2.36, has no subgroup of order 6, even though $6 \mid 12 = |A_4|$. Cauchy's Theorem ensures the existence of cyclic subgroups of order 2 and 3, because these are the prime factors of $|A_4| = 12 = 2^2 \cdot 3$. In fact, there are six subgroups of order 2:

$$\begin{array}{lll} \{\iota, (12)\}, & \{\iota, (13)\}, & \{\iota, (14)\}, \\ \{\iota, (23)\}, & \{\iota, (24)\}, & \{\iota, (34)\}. \end{array}$$

There are four subgroups of order 3:

$$\begin{array}{ll} \{\iota, (123), (132)\}, & \{\iota, (124), (142)\}, \\ \{\iota, (134), (143)\}, & \{\iota, (234), (243)\}. \end{array}$$

There is also one subgroup of order 4:

$$\{\iota, (12)(34), (13)(24), (14)(23)\}.$$

What patterns can we find here? The subgroup of order 4 isn't predicted by Cauchy's Theorem, but it exists nonetheless. Meanwhile, we know from Example 2.36 that no subgroup of order 6 exists. What property does 4 have that 6 doesn't? Well, neither are prime, but $4 = 2^2$ is at least a power of a prime. This, it turns out, is significant. To explore this idea further, we need to introduce another bit of terminology.

Definition 7.1 Let $p \in \mathbb{N}$ be prime, and let G be a group of order $|G| = p^n$ for some integer $n \geq 0$. Then we say G is a **p -group**.

An immediate consequence of this definition and Proposition 2.34 is that every element of a p -group must have order a power of p . Indeed, we could happily have defined p -groups as groups for which every element has order p^k for some $k \geq 0$, and some books do just that.

There's another related concept that we'll use in this section:

Definition 7.2 Let G be a group, and let $H \leq G$ be a subgroup with $|H| = p^n$ for some prime $p \in \mathbb{N}$ and some integer $n \geq 0$. Then H is a **p -subgroup** of G ; that is, a subgroup of G that is itself a p -group.

As we've seen, the alternating group A_4 doesn't have subgroups of all possible orders, but it does have p -subgroups of all possible orders.

The next smallest counterexample to the converse of Lagrange's Theorem is $SL_2(3)$, the group of 2×2 matrices with elements from $\mathbb{F}_3$ and determinant 1. This has order $24 = 2^3 \cdot 3$, and hence by Cauchy's Theorem should have at least one cyclic subgroup of order 2 and

another of order 3. In fact, it has one subgroup

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \right\} \cong \mathbb{Z}_2$$

of order 2, three subgroups

$$\langle \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 1 & 2 \\ 2 & 2 \end{bmatrix} \rangle \cong \mathbb{Z}_4$$

of order 4, one subgroup

$$\langle \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 2 & 2 \\ 2 & 1 \end{bmatrix} \rangle \cong Q_8$$

of order 8, four subgroups

$$\langle \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 0 & 1 \\ 2 & 2 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix} \rangle \cong \mathbb{Z}_3$$

of order 3, and four subgroups

$$\langle \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 0 & 1 \\ 2 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \rangle, \quad \langle \begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix}, \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \rangle \cong \mathbb{Z}_6$$

of order 6. By Lagrange's Theorem, it could also have a subgroup of order 12, but it turns out not to. Again, however, it does have p -subgroups of all possible orders.

This happens to be true in general, and to prove it we first need to take a little detour into the realm of group actions. Let G be a (possibly infinite) group acting on a finite set X . Then there are finitely many orbits in X under this action, and these orbits partition X :

$$|X| = \sum_{i=1}^r |\text{Orb}_G(x_i)| \quad (7.1)$$

where $x_1, \dots, x_r$ are representative elements from each orbit. Denote by $\text{Fix}_G(X)$ the set of elements of X that are fixed by every element of G . Each of these elements must therefore be a single-element orbit of the G -action, so $\text{Fix}_G(X)$ is the union of these singleton orbits. Suppose there are s of these, with $0 \leq s \leq r$. Then $|\text{Fix}_G(X)| = s$ and we can rewrite (7.1) as

$$|X| = |\text{Fix}_G(X)| + \sum_{i=s+1}^r |\text{Orb}_G(x_i)| \quad (7.2)$$

In the special case where G is a p -group, we obtain the following useful fact.

Proposition 7.3 *Let G be a p -group of order p^k for some prime p and positive integer k , and suppose that G acts on some finite set X . Then*

$$|X| \equiv |\text{Fix}_G(X)| \pmod{p}.$$

Proof By the Orbit-Stabiliser Theorem⁶ we know that $|\text{Orb}_G(x)|$ is a factor of $|G|$ for any $x \in X$. Hence in (7.2) it follows that p divides $|\text{Orb}_G(x_i)|$ for $s+1 \leq i \leq r$. Hence

$$|X| = |\text{Fix}_G(X)| + \sum_{i=s+1}^r |\text{Orb}_G(x_i)| \equiv |\text{Fix}_G(X)| \pmod{p}$$

as claimed. □

⁶ Theorem 6.25, page 217.

What this tells us is that the modulo- p congruence class of the number of elements in X is determined solely by the single-element orbits; that is, the points that are fixed by every element of G .

We are now ready to prove the fact we mentioned earlier: that a finite group G has p -subgroups of all possible orders for any prime factor p of $|G|$. We will call this Sylow's First Theorem.⁷

⁷This is actually a slightly stronger version of what most books call Sylow's First Theorem. Also, there isn't a complete consensus on the numbering of Sylow's Theorems: some books combine them into one large theorem with several parts, while others divide them into three or sometimes four separate theorems; neither is there complete agreement on the order in which they should be presented. Here we will state and prove them as three separate but closely-related theorems, in what seems these days to be the most common order.

⁸Theorem 2.37, page 57.

Theorem 7.4 (Sylow's First Theorem) *Let G be a finite group of order $|G| = p^k m$, where $p \in \mathbb{Z}$ is prime, $k \in \mathbb{Z}$ is non-negative, and $m \in \mathbb{N}$ with $p \nmid m$. Then G contains a subgroup of order p^i for all $0 \leq i \leq k$, and each subgroup of order p^i is a normal subgroup of some subgroup of order p^{i+1} for $0 \leq i < k$.*

Proof We will proceed by induction on i . The case $i = 0$ is trivial, and the case $i = 1$ holds by Cauchy's Theorem⁸. Suppose, then, that H is a p -subgroup of G of order $|H| = p^i$ for some $i \geq 1$. Let S be the set of left cosets of H in G , and define a left action of H on S by

$$h \cdot (gH) = (hg)H.$$

By Proposition 7.3 the number of cosets in S is congruent modulo p to the number that are fixed by the action of H :

$$|S| \equiv |\text{Fix}_H(S)| \pmod{p}.$$

For a coset gH to be fixed by this action means that $h \cdot (gH) = (hg)H = gH$ for all $h \in H$. Proposition 2.26 tells us that this is equivalent to saying that $hg \in gH$ for all $h \in H$, which is the same as requiring that $g^{-1}hg \in H$ for all $h \in H$. This, in turn, is equivalent to saying that g lies in the normaliser $N_G(H)$ of H in G .

So, the cosets gH fixed by this action are exactly those for which $g \in N_G(H)$, and hence $\text{Fix}_H(S) = \{gH : g \in N_G(H)\}$ we find that

$$|\text{Fix}_H(S)| = |N_G(H):H|.$$

Looking at this another way, the left cosets of H in G are all the same size and partition G , so $|S| = |G:H|$. Putting all this together we get

$$|G:H| = |S| \equiv |\text{Fix}_H(S)| = |N_G(H):H| \pmod{p}. \quad (7.3)$$

Recall⁹ that the normaliser $N_G(H)$ is the largest subgroup of G in which H is normal, so $H \trianglelefteq N_G(H)$. Therefore it makes sense to talk about the quotient group $N_G(H)/H$.

Also, H is a p -subgroup of G , so $|H| = p^i$ for some $i \geq 1$. So if $i < k$ the index $|G:H|$ is a multiple of p , and by (7.3) it follows that $|N_G(H):H|$ is also a multiple of p . Thus $|N_G(H)/H|$ is a multiple of p too, and by Cauchy's Theorem $N_G(H)/H$ has a subgroup of order p .

All subgroups of $N_G(H)/H$ are of the form K/H for some subgroup K such that $H \leq K \leq N_G(H)$. So there exists some subgroup K/H

⁹Definition 4.46, page 120.

in $N_G(H)/H$ of order p , which means that $|K:H| = p$. Hence $|K| = p|H| = p^{i+1}$. This completes the inductive step, and hence G contains p -subgroups of all possible orders.

Finally, as H is normal in $N_G(H)$ it must be normal in $K \leq N_G(H)$. $\square$

In many cases, therefore, Sylow's First Theorem gives us more information about orders of subgroups than Cauchy's Theorem.

Example 7.5 Suppose that G is a group of order $12 = 2^2 \cdot 3$. Then Cauchy's Theorem tells us that G has (cyclic) subgroups of orders 2 and 3, because these are the prime factors of 12. But Sylow's First Theorem tells us that G must have a subgroup of order $4 = 2^2$ as well. Not only that, each subgroup of order 2 is contained (and is normal) in a subgroup of order 4.

Example 7.6 Let G be a group of order $8 = 2^3$. We know from Proposition 2.40 that, up to isomorphism, there are five different such groups. Cauchy's Theorem tells us that G must have at least one subgroup of order 2, but Sylow's First Theorem also ensures the existence of subgroups of order 4 as well. Also, these order-4 subgroups must be normal in G .

Of course, we know this last bit of information already because by Proposition 3.12 any index-2 subgroup must be normal. If, however, $|G| = 27 = 3^3$ then G must have subgroups of order 3 and 9, and the order-9 subgroups must all be normal.

We want to know more about these subgroups: how are they related and how many are there? Sylow's other theorems give answers to these questions, so we'll look carefully at some small examples.

Example 7.7 The dihedral group D_3 has order $6 = 2 \cdot 3$, and hence by Cauchy's Theorem (and also by Sylow's First Theorem) it must have at least one subgroup of order 2 and at least one of order 3. In fact, we know from earlier discussions that it has three subgroups

$$M_1 = \{e, m_1\}, \quad M_2 = \{e, m_2\} \quad \text{and} \quad M_3 = \{e, m_3\}$$

of order 2. These represent the reflections in the three axes of the equilateral triangle. We might ask how these are related, and the answer is given by Example 3.5: the three reflections m_1 , m_2 and m_3 are all conjugate to each other, and the identity element e forms a singleton conjugacy class on its own.

Upon further inspection, then, we can reconstruct the subgroup $M_2 = \{e, m_2\}$ from the subgroup $M_1 = \{e, m_1\}$ by conjugating each element by the $\frac{2\pi}{3}$ rotation r :

$$e = rer^{-1}, \quad m_2 = rm_1r^{-1}.$$

Similarly, we can reconstruct the subgroup $M_3 = \{e, m_3\}$ by conjugating with the $\frac{4\pi}{3}$ rotation $r^2 = r^{-1}$:

$$e = r^{-1}er, \quad m_3 = r^{-1}m_1r.$$

So, these subgroups are conjugate to each other:

$$M_2 = rM_1r^{-1} \quad \text{and} \quad M_3 = r^{-1}M_1r.$$

There is only one subgroup of order 3: the rotation subgroup $R_3 = \{e, r, r^2\}$. This is a normal subgroup and is hence conjugate to itself.

Let's look at a slightly larger example: the symmetric group S_4 .

Example 7.8 The symmetric group S_4 has order $24 = 2^2 \cdot 3$, and hence by Sylow's First Theorem it must have at least one subgroup each of orders 2, 3, 4 and 8. In fact, it has three subgroups of order 8, each isomorphic to the dihedral group D_4 :

$$\langle (1234), (13) \rangle, \quad \langle (1243), (14) \rangle \quad \text{and} \quad \langle (1324), (12) \rangle.$$

These are all conjugate to each other; to see this, we only have to check the two generators for each subgroup:

$$\begin{aligned} (34)(1234)(34) &= (1243), & (34)(13)(34) &= (14), \\ (23)(1234)(23) &= (1324), & (23)(13)(23) &= (12). \end{aligned}$$

There should also be at least one subgroup of order 3; in fact there are four:

$$\langle (123) \rangle, \quad \langle (124) \rangle, \quad \langle (134) \rangle \quad \text{and} \quad \langle (234) \rangle.$$

These are also conjugate to each other:

$$\begin{aligned} (34)(123)(34) &= (124), & (23)(124)(23) &= (134), \\ (12)(134)(12) &= (234), & (14)(234)(14) &= (123). \end{aligned}$$

The pattern we're starting to see here is that if p is a prime factor of $|G|$, then the maximal p -subgroups of G are conjugate to each other. Does this happen for smaller p -subgroups?

Example 7.9 The group S_4 has seven subgroups of order $4 = 2^2$:

$$\begin{aligned} &\langle (1234) \rangle, \quad \langle (1243) \rangle, \quad \langle (1324) \rangle, \\ &\langle (13), (24) \rangle, \quad \langle (14), (23) \rangle, \quad \langle (12), (34) \rangle, \\ &\langle (12)(34), (13)(24) \rangle \end{aligned}$$

The first three of these are isomorphic to the cyclic group $\mathbb{Z}_4$, while the remaining four are isomorphic to the Klein group $V \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. But conjugation by an element $g \in G$ yields isomorphisms $f_g: H \rightarrow K$, where $f_g(h) = ghg^{-1}$ for all $h \in H$, from one subgroup H to some other subgroup K . And since not all of the order-4 subgroups aren't

isomorphic, at least some of them can't be conjugate to each other.

So this pattern only seems to work for maximal p -subgroups. Subgroups of this type are named after Ludwig Sylow himself:

Definition 7.10 Let G be a finite group of order $|G| = p^k m$ for some prime $p \in \mathbb{N}$, some integer $n \geq 0$ and some $m \in \mathbb{N}$ such that $p \nmid m$. Then a p -subgroup of order p^k is said to be a **Sylow p -subgroup**.

The pattern we see from the above examples is that for any prime factor p of $|G|$, the Sylow p -subgroups are conjugate to each other. This is true in general, a result that we will call Sylow's Second Theorem:

Theorem 7.11 (Sylow's Second Theorem) *Let G be a finite group with order $|G| = p^k m$ for some prime p , non-negative integer k and positive integer m not divisible by p . If H and K are Sylow p -subgroups of G , then H and K are conjugate. That is, there exists some element $g \in G$ such that $gHg^{-1} = K$.*

Proof If $|G| = p^k$ then G is itself the unique Sylow p -subgroup, is obviously closed under conjugation, and the result follows immediately. Suppose otherwise, that G is not a p -group, and let H and K be two Sylow p -subgroups of G , with $|H| = |K| = p^k$. Let S be the set of left cosets of H in G , and let K act on S by left multiplication:

$$k \cdot (gH) = (kg)H$$

for all $k \in K$ and $g \in G$. We know that K is a finite p -group, so by (7.2) we have

$$|S| \equiv |\text{Fix}_K(S)| \pmod{p}.$$

Observe that $|S| = |G:H| = |G|/|H| = m$. By the hypothesis, $p \nmid m$ and so $|S| \not\equiv 0 \pmod{p}$. Therefore $|\text{Fix}_K(S)| \neq 0$ and so there is at least one left coset gH in S that is fixed by the K -action. This means that $(kg)H = gH$ for all $k \in K$. By Proposition 2.26 this is equivalent to saying that $kg \in gH$ for all $k \in K$, and hence $K \subseteq gHg^{-1}$. But $|K| = |gHg^{-1}|$ and hence $K = gHg^{-1}$, so K and H are conjugate. $\square$

So all the Sylow p -subgroups of a finite group G are conjugate to each other. If G has only one Sylow p -subgroup, then it is mapped to itself by conjugation by every element of G , and is therefore normal. We saw this in Example 7.7: D_3 has only one Sylow 3-subgroup, the rotation subgroup R_3 , which we already know is normal. But this happens to be true in general:

Corollary 7.12 *Let G be a finite group, and let p be a prime factor of $|G|$. Then G has exactly one Sylow p -subgroup H if and only if H is normal.*

This leads us neatly to the second of our questions about Sylow p -subgroups: how many are there?

G	$ G $	p	n_p	$ G:H $
D_3	6	2	3	3
		3	1	2
A_4	12	2	1	3
		3	4	4
$\mathbb{Z}_{12}$	12	2	1	3
		3	1	4
S_4	24	2	3	3
		3	4	8
$SL_2(3)$	24	2	1	3
		3	4	8
$\mathbb{Z}_{10}$	10	2	1	5
		5	1	2
D_5	10	2	5	5
		5	1	2

Table 7.1: The number of Sylow p -subgroups H contained in a group G

In order to explore this question, we'll introduce a couple of useful bits of notation. Suppose that G is a finite group, and p is a prime integer that divides $|G|$. Let $\text{Syl}_p(G)$ denote the set of Sylow p -subgroups contained in G , and let $n_p = |\text{Syl}_p(G)|$.

For example, if $G = D_3$ then $n_2 = 3$ and $n_3 = 1$ since D_3 has three Sylow 2-subgroups and one Sylow 3-subgroup. Table 7.1 lists some examples of small finite groups.

What patterns can we see in these examples? All of them have an odd number of Sylow 2-subgroups. The first five examples all have either one or four Sylow 3-subgroups, and the last two examples both have a single Sylow 5-subgroup.

The key observation here is to count modulo p . Given that crucial steps in the proofs of Sylow's First and Second Theorems required us to count cosets modulo p , in retrospect this shouldn't be a colossal surprise. Counting modulo p is apparently a fundamental aspect of working with p -subgroups. Looking at these examples through a modulo- p lens, we find that in every case the number n_p of Sylow p -subgroups is congruent to 1.

Also, in each case n_p is a factor of the order of the group in question. But more than that, it is a factor of the index of the given Sylow p -subgroup. Both of these are true in general, by a result we will call Sylow's Third Theorem.

Theorem 7.13 (Sylow's Third Theorem) *Let G be a finite group of order $|G| = p^k m$ where p is prime, k is a non-negative integer, and m is a positive integer not divisible by p . Then $n_p \equiv 1 \pmod{p}$ and $n_p | m$.*

Proof As with Sylow's first two Theorems, we will study the action of a particular group on a suitably chosen set, and then apply (7.2).

To prove the first statement, suppose H is a Sylow p -subgroup of G and let it act on $\text{Syl}_p(G)$ by conjugation. For any Sylow p -subgroup K to be fixed under this action, we require $hKh^{-1} = K$ for all $h \in H$; the subgroup H itself is certainly fixed in this way.

Suppose that K is any Sylow p -subgroup fixed under conjugation by H . Then H is contained in the normaliser $N_G(K)$ of K in G ; clearly also $K \subseteq N_G(K)$, so both H and K are Sylow p -subgroups in $N_G(K)$, and by Sylow's Second Theorem they must be conjugate in $N_G(K)$.

But K is normal in, and hence conjugate to itself in, its own normaliser $N_G(K)$. So $K = H$ and thus the only Sylow p -subgroup in $\text{Syl}_p(G)$ fixed by conjugation with H is H itself. Hence $|\text{Fix}_H(\text{Syl}_p(G))| = 1$ and by (7.2) we have, as required,

$$n_p = |\text{Syl}_p(G)| \equiv |\text{Fix}_H(\text{Syl}_p(G))| = 1 \pmod{p}.$$

To show that $n_p | m$ we let the full group G act on $\text{Syl}_p(G)$ by conjugation. By Sylow's Second Theorem, all the Sylow p -subgroups of G are conjugate to each other, so the conjugation action has only a single orbit, namely the whole of $\text{Syl}_p(G)$. By the Orbit-Stabiliser Theorem,¹⁰ the number of elements in an orbit divides the order of the group, so $n_p = |\text{Syl}_p(G)|$ must divide $|G| = p^k m$. Since we've just shown that $n_p \equiv 1 \pmod{p}$, and hence n_p is coprime to p , the only remaining possibility is that $n_p | m$, as claimed. $\square$

¹⁰ Theorem 6.25, page 217.

We can use Sylow's Theorems to help us classify finite groups of a given order, as the next example shows.

Example 7.14 Let G be a group of order $|G| = 15 = 3 \cdot 5$. Sylow's First Theorem tells us that G must contain subgroups of order 3 and 5. (Actually, we know this from Cauchy's Theorem too.) In particular, these subgroups are Sylow 3- and 5-subgroups.

Sylow's Third Theorem says that $n_3 \equiv 1 \pmod{3}$ and $n_3 | 5$; the first of these means that n_3 could be 1, 4, 7, 10 or 13, and the second (or, alternatively, Lagrange's Theorem) rules out all but $n_3 = 1$. So there is one Sylow 3-subgroup, isomorphic to $\mathbb{Z}_3$.

Considering the Sylow 5-subgroups in the same way, we find the $n_5 \equiv 1 \pmod{5}$ and $n_5 | 3$, so n_5 must also equal 1. Thus G has a single Sylow 5-subgroup, isomorphic to $\mathbb{Z}_5$.

Sylow's Second Theorem says that Sylow p -subgroups are conjugate to each other, and since $n_3 = n_5 = 1$, both the order-3 and order-5 subgroups are normal in G . In particular, G can't be simple.

Let H be the order-3 subgroup and K be the order-5 subgroup. Their intersection $H \cap K$ must be a subgroup of both H and K , and Lagrange's Theorem says that $|H \cap K|$ must therefore divide both $|H| = 3$ and $|K| = 5$. Since 3 and 5 are coprime, the only possibility is that $|H \cap K| = 1$, and so $H \cap K = \{e\}$. The subgroup HK has order greater than 5, and hence by Lagrange's Theorem, $|HK|$ must equal 15, so HK must be the full group G .

By Corollary 3.23, $G \cong H \times K$, and therefore $G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_5$. Finally, since 3 and 5 are coprime, Proposition 1.32 tells us that $G \cong \mathbb{Z}_{15}$.

Up to isomorphism, then, there is only one group of order 15: the cyclic group $\mathbb{Z}_{15}$.

More generally, we can prove the following:

Proposition 7.15 *Let G be a group of order pq where p and q are both prime, $p < q$ and $q \not\equiv 1 \pmod{p}$. Then $G \cong \mathbb{Z}_{pq}$.*

Proof By Lagrange's Theorem, any proper nontrivial subgroup of G must have order either p or q . By Sylow's First Theorem, or Cauchy's

Theorem, G must have at least one subgroup isomorphic to $\mathbb{Z}_p$ and at least one subgroup isomorphic to $\mathbb{Z}_q$.

Sylow's Third Theorem tells us that $n_p \equiv 1 \pmod{p}$ and $n_p | q$, so either $n_p = 1$ or q , but the latter case is disallowed by the hypothesis $q \not\equiv 1 \pmod{p}$, hence $n_p = 1$. Similarly, $n_q \equiv 1 \pmod{q}$ and $n_q | p$, so either $n_q = 1$ or p , but the second of these can't happen because $p < q$, so $p \not\equiv 1 \pmod{q}$, and hence $n_q = 1$.

Let H be the subgroup of order p and K be the subgroup of order q . Then by Corollary 7.12 these are both normal in G . Their intersection $H \cap K$ is a subgroup of G and also of both H and K by Proposition 2.9. By Proposition 2.11, HK is a subgroup of G and its order must be greater than $|K| = q$, so again by Lagrange's Theorem the only possibility is that $|HK| = |G| = pq$, so $HK = G$.

By Corollary 3.23, then, $G \cong H \times K \cong \mathbb{Z}_p \oplus \mathbb{Z}_q$. And since p and q are coprime, Proposition 1.32 tells us that $G \cong \mathbb{Z}_{pq}$ and hence any group of order pq is cyclic. $\square$

Corollary 7.16 *Let p and q be prime, with $p < q$ and $q \not\equiv 1 \pmod{p}$. Then any group of order pq is abelian.*

We noted earlier that, up to isomorphism, there are fifteen groups of order 24.¹¹ We'll leave the details of the classification for Section 7.A, but we can use Sylow's Theorems to find out what p -subgroups exist:

Example 7.17 Since $24 = 2^3 \cdot 3$, Sylow's First Theorem guarantees the existence of subgroups of order 2, 3, 4 and 8.

By Sylow's Third Theorem, $n_2 \equiv 1 \pmod{2}$ and $n_2 | 3$, so there are either 1 or 3 Sylow 2-subgroups of order 8. Sylow's Second Theorem says that these are conjugate to each other.

Similarly, $n_3 \equiv 1 \pmod{3}$ and $n_3 | 8$. The first of these allows n_3 to be 1, 4, 7, 10, 13, 16, 19 and 22, but the second (and Lagrange's Theorem) rules out all but $n_3 = 1$ or 4.

So, any group of order 24 must have at least one subgroup of orders 2 and 4, and either 1 or 3 subgroups of order 8 and 1 or 4 subgroups of order 3.

We can use Corollary 7.12 to show that there are no simple groups of certain orders:

Example 7.18 Let G be a group of order $|G| = 1246 = 2 \cdot 7 \cdot 89$. Then by Sylow's First Theorem (or Cauchy's Theorem) G has (Sylow) subgroups of order 2, 7 and 89. Sylow's Third Theorem says that n_2 is odd and divides $7 \cdot 89 = 623$. Therefore G has either 1, 7, 89 or 623 subgroups of order 2.

Similarly, Sylow's Third Theorem tells us that $n_7 \equiv 1 \pmod{7}$ and

¹¹ Table 4.1, page 119.

$n_7|178$, whence $n_7 = 1$, and that $n_{89} \equiv 1 \pmod{89}$ and $n_{89}|14$, so $n_{89} = 1$.

By Corollary 7.12, the subgroup of order 7 and the subgroup of order 89 must be normal in G , and so no group of order 1246 can be simple.

Corollary 7.16 illustrated another useful application of Sylow's Theorems: to show that groups of certain orders are abelian. We'll take this idea a little further now. To start with, we'll use (7.2) to show that p -groups have nontrivial centres, and then we'll use that to classify groups of order p^2 and show that they are all abelian.

Proposition 7.19 *Let G be a finite p -group of order p^k , for some prime p and integer $k > 0$. Then the centre $Z(G)$ is nontrivial.*

Proof Let G act on itself by conjugation. Then the set $\text{Fix}_G(G)$ consists of all elements $h \in G$ for which $ghg^{-1} = h$ for all $g \in G$. This is equivalent to requiring $gh = hg$ for all $g \in G$, hence $\text{Fix}_G(G) = Z(G)$.

Applying 7.2, we get

$$|G| \equiv |\text{Fix}_G(G)| = |Z(G)| \pmod{p}.$$

We know that $|G|$ is a power of p , so $|Z(G)| \equiv 0 \pmod{p}$ and hence $|Z(G)|$ is a multiple of p . But $Z(G)$ is nonempty: at the very least it must contain the identity. And since $p > 1$ it follows that $Z(G)$ must contain at least $(p-1)$ other elements apart from the identity. $\square$

Now we can prove that groups of order p^2 are abelian.

Proposition 7.20 *Let G be a group of order p^2 for some prime p . Then G is isomorphic either to $\mathbb{Z}_{p^2}$ or $\mathbb{Z}_p \oplus \mathbb{Z}_p$.*

Proof By Proposition 2.34, the order of any non-identity element of G must be either p or p^2 . If G has an element g of order p^2 then $G = \langle g \rangle \cong \mathbb{Z}_{p^2}$.

Otherwise, every non-identity element g has order p . By Proposition 7.19 at least one of these lies in the centre $Z(G)$ and hence commutes with every element of G . Choose one such non-identity element $g \in Z(G)$. Then $\langle g \rangle$ is a cyclic subgroup of order p , isomorphic to $\mathbb{Z}_p$. Now choose some element $h \in G \setminus \langle g \rangle$; this also has order p .

Let $f: \mathbb{Z}_p \oplus \mathbb{Z}_p \rightarrow G; (m, n) \mapsto g^m h^n$. This is a homomorphism, since

$$\begin{aligned} f(m_1, n_1)f(m_2, n_2) &= (g^{m_1} h^{n_1})(g^{m_2} h^{n_2}) \\ &= g^{m_1+m_2} h^{n_1+n_2} = f(m_1+m_2, n_1+n_2) \end{aligned}$$

and since g and all its powers are in $Z(G)$, so they commute with everything, in particular any power of h .

The kernel $\ker(f)$ is trivial: if $f(m, n) = e$ then $g^m h^n = e$ and so $g^m = h^{-n}$. This element must lie in the intersection $\langle g \rangle \cap \langle h \rangle$, and the only such element is the identity $e \in G$. So $g^m = h^n = e$ and thus $m = n = 0$ in $\mathbb{Z}_p$. Since $\ker(f)$ is trivial, f is injective. And since $|G| = |\mathbb{Z}_p \oplus \mathbb{Z}_p| = p^2$, then f is surjective. Hence $G \cong \mathbb{Z}_p \oplus \mathbb{Z}_p$. $\square$

Because cyclic groups and direct sums of cyclic groups are abelian, we have the following corollary.

Corollary 7.21 *Any group G of order $|G| = p^2$ is abelian.*

The next proposition takes this idea a little further.

Proposition 7.22 *Let p and q be prime, with $p < q$ and $q \not\equiv 1 \pmod{p}$. Then any group of order $p^2 q$ is abelian.*

Proof A Sylow p -subgroup of G has order p^2 , while a Sylow q -subgroup has order q . By Sylow's Third Theorem, $n_p | q$ and $n_p \equiv 1 \pmod{p}$, while $n_q | p$ and $n_q \equiv 1 \pmod{q}$. The only possibility is that $n_p = n_q = 1$.

Let H be the Sylow p -subgroup and K be the Sylow q -subgroup. Both are normal in G by Corollary 7.12.

By Proposition 7.20, H is isomorphic either to $\mathbb{Z}_{p^2}$ or $\mathbb{Z}_p \oplus \mathbb{Z}_p$, and by Corollary 7.21 is therefore abelian. The group K is isomorphic to $\mathbb{Z}_q$ and is hence also abelian. Furthermore, their intersection is trivial because p^2 and q are coprime, and therefore by Propositions 3.22 and 2.11 HK is a subgroup of G . $|HK| = p^2 q = |G|$, so $HK = G$. By Corollary 3.23, $G \cong H \times K$ so G must be abelian. Hence either $G \cong \mathbb{Z}_{p^2} \oplus \mathbb{Z}_q$ or $G \cong \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_q$. $\square$

For example, we now know that any group of order $45 = 3^2 \cdot 5$ must be isomorphic either to $\mathbb{Z}_9 \oplus \mathbb{Z}_5$ or $\mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_5$.

If the material world rests upon a similar ideal world, this ideal world must rest upon some other; and so on, without end. It were better, therefore, never to look beyond the present material world.

— David Hume (1711–1776),
Dialogues Concerning Natural Religion (1779)

7.2 Series of subgroups

IN CHAPTER 2 we met the concept of a subgroup lattice or Hasse diagram of a group G : a directed graph whose nodes correspond to the various subgroups of G and whose edges encode the relationship “is a subgroup of”. For example, the Klein group $V \cong \mathbb{Z} \oplus \mathbb{Z}$ has five subgroups in total: itself, three subgroups isomorphic to $\mathbb{Z}_2$, and the trivial subgroup. The corresponding graph is shown in Figure 7.1. This is helpful for picturing the subgroup structure of a fairly small group, but it rapidly becomes unwieldy for larger groups.

But we can still learn a lot from studying how subgroups fit together. If

we take a subgroup lattice for a finite group G , then any path through the graph starting at G and ending at the trivial subgroup gives a nested series of subgroups, each one containing the next.

For example, one such series for the Klein group is

$$\{e\} < \{e, a\} < V,$$

while for the dihedral group D_4 we have

$$\{e\} < \{e, r^2\} < \{e, r, r^2, r^3\} < D_4$$

and

$$\{e\} < \{e, m_1\} < \{e, r^2, m_1, m_3\} < D_4$$

amongst others. We call these chains of nested subgroups **series**.¹²

Definition 7.23 A **series** (or **subgroup series**) for a group G is a finite sequence of subgroups

$$\{e\} = H_0 < H_1 < \cdots < H_n = G$$

of G such that H_i is a proper subgroup of H_{i+1} for $0 \leq i < n$. The **length** of the series is the number n of inclusions $H_i < H_{i+1}$.

Sometimes it will be convenient to number the subgroup indices in the reverse order. We will call a series of the form

$$H_0 < H_1 < \cdots < H_n = G$$

an **ascending series** for G , and a series of the form

$$H_n < H_{n-1} < \cdots < H_0 = G$$

a **descending series** for G .

We may also sometimes relax the requirement that the smallest subgroup (H_0 for an ascending series, H_n for a descending series) is trivial.

We've seen already that normal subgroups have nice and useful properties, and it turns out that for our purposes it will be most useful to focus on series consisting of subgroups of this type. But there are a couple of different ways of doing this: we can either insist that all the subgroups $H_0, \dots, H_n$ are normal in G , or we can just require that each subgroup H_i is normal in the next one along, H_{i+1} .

Definition 7.24 A series

$$H_0 = \{e\} < H_1 < \cdots < H_{n-1} < H_n = G$$

is **normal** or **invariant** if $H_0, \dots, H_n \trianglelefteq G$; that is, if each subgroup H_i is normal in G .

A series is **subnormal** or **subinvariant** if $H_i \triangleleft H_{i+1}$ for $0 \leq i < n$; that is, if each subgroup is normal in the next one in the list.

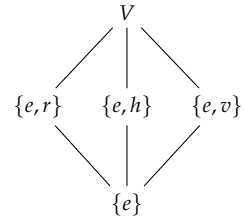


Figure 7.1: Subgroup lattice for the Klein group V

¹² This conflicts with the usage in analysis, where a sequence is a discrete list and a series is the sum of an infinite sequence.

Any group G has a normal series of length 1, namely $\{e\} \triangleleft G$. If G is simple, then this is the only normal series it can have. If G is abelian, then any series is both normal and subnormal.

Every normal series is subnormal, but not every subnormal series is normal. For example, the series

$$\{e\} \triangleleft \{e, m_1\} \triangleleft \{e, r^2, m_1, m_3\} \triangleleft D_4 \quad (7.4)$$

is subnormal, but not normal, because $\{e, m_1\} \not\triangleleft D_4$.

¹³ Theorem 7.4, page 236.

Sylow's First Theorem¹³ guarantees the existence of subnormal series of p -subgroups for any finite group G .

The point of insisting on normal subgroups is that we can then pass to quotient groups, secure in the knowledge that each quotient H_{i+1}/H_i is well-defined. For example, with the subnormal series for D_4 shown above, we get the quotients

$$H_1/H_0 \cong \mathbb{Z}_2, \quad H_2/H_1 \cong \mathbb{Z}_2 \quad \text{and} \quad H_3/H_2 \cong \mathbb{Z}_2.$$

There are other subnormal series for D_4 , for example

$$\{e\} \triangleleft \{e, m_2\} \triangleleft \{e, r^2, m_2, m_4\} \triangleleft D_4, \quad (7.5)$$

the quotients for which are

$$H_1/H_0 \cong \mathbb{Z}_2, \quad H_2/H_1 \cong \mathbb{Z}_2 \quad \text{and} \quad H_3/H_2 \cong \mathbb{Z}_2.$$

Although the series (7.4) and (7.5) aren't the same, the corresponding collection of quotient groups are isomorphic.

The group $\mathbb{Z}_6$ has subgroups $\langle 2 \rangle = \{0, 2, 4\}$ and $\langle 3 \rangle = \{0, 3\}$, which both yield subnormal (and, since $\mathbb{Z}_6$ is abelian, normal) series for $\mathbb{Z}_6$:

$$\{0\} \triangleleft \langle 2 \rangle \triangleleft \mathbb{Z}_6 \quad (7.6)$$

$$\{0\} \triangleleft \langle 3 \rangle \triangleleft \mathbb{Z}_6 \quad (7.7)$$

The quotient groups obtained from the first of these series are

$$H_1/H_0 \cong \mathbb{Z}_3 \quad \text{and} \quad H_2/H_1 \cong \mathbb{Z}_2,$$

while those obtained from the second series are

$$H_1/H_0 \cong \mathbb{Z}_2 \quad \text{and} \quad H_2/H_1 \cong \mathbb{Z}_3.$$

Again, these two series are different, but their quotient groups are the same, at least up to permutation. This turns out to be a useful way of comparing normal and subnormal series:

Definition 7.25 Two normal or subnormal series

$$\{e\} = H_0 \triangleleft \cdots \triangleleft H_n = G \quad \text{and} \quad \{e\} = K_0 \triangleleft \cdots \triangleleft K_n = G$$

are **isomorphic** or **equivalent** if there is a bijection between the sets

$$\{H_{i+1}/H_i : 0 \leq i < n\} \quad \text{and} \quad \{K_{i+1}/K_i : 0 \leq i < n\}.$$

By this definition, both subnormal series (7.4) and (7.5) for D_4 are isomorphic, and both normal series (7.6) and (7.7) for $\mathbb{Z}_6$ are isomorphic too. We can take this idea a little further. Consider the normal (and subnormal) series

$$\{0\} \triangleleft 60\mathbb{Z} \triangleleft 12\mathbb{Z} \triangleleft 2\mathbb{Z} \triangleleft \mathbb{Z} \quad (7.8)$$

for $\mathbb{Z}$. There is still room to fit more subgroups in, and we can do this in different ways. For example,

$$\{0\} \triangleleft 120\mathbb{Z} \triangleleft 60\mathbb{Z} \triangleleft 12\mathbb{Z} \triangleleft 6\mathbb{Z} \triangleleft 2\mathbb{Z} \triangleleft \mathbb{Z} \quad (7.9)$$

and

$$\{0\} \triangleleft 180\mathbb{Z} \triangleleft 60\mathbb{Z} \triangleleft 12\mathbb{Z} \triangleleft 4\mathbb{Z} \triangleleft 2\mathbb{Z} \triangleleft \mathbb{Z} \quad (7.10)$$

are both valid normal series for $\mathbb{Z}$, obtained by fitting extra groups into the series (7.8). They aren't isomorphic: the quotient groups of (7.9) are

$$120\mathbb{Z} \cong \mathbb{Z}, \quad \mathbb{Z}_2, \quad \mathbb{Z}_5, \quad \mathbb{Z}_2, \quad \mathbb{Z}_3 \quad \text{and} \quad \mathbb{Z}_2,$$

while the quotient groups of (7.10) are

$$180\mathbb{Z} \cong \mathbb{Z}, \quad \mathbb{Z}_3, \quad \mathbb{Z}_5, \quad \mathbb{Z}_3, \quad \mathbb{Z}_2 \quad \text{and} \quad \mathbb{Z}_2.$$

These series aren't isomorphic to each other, and neither are they isomorphic to the original series (7.8), but this is a useful idea that's worth studying further, and to that end we introduce the following definition.

Definition 7.26 Let $H = \{H_0, \dots, H_m\}$ and $\{K_0, \dots, K_n\}$ be two normal or subnormal series of a group G . Then we say that K is a **refinement** of H if we can obtain K by inserting additional subgroups into H . More precisely, there exists an injective, strictly increasing function

$$f: \{0, \dots, m\} \hookrightarrow \{0, \dots, n\}$$

such that $H_i = K_{f(i)}$ for $0 \leq i \leq m$.

Thus the series (7.9) and (7.10) are (non-isomorphic) refinements of the series (7.8). Now consider the series

$$\{0\} \triangleleft 3\mathbb{Z} \triangleleft \mathbb{Z} \quad (7.11)$$

with quotient groups

$$3\mathbb{Z} \cong \mathbb{Z} \quad \text{and} \quad \mathbb{Z}_3,$$

and the series

$$\{0\} \triangleleft 8\mathbb{Z} \triangleleft 2\mathbb{Z} \triangleleft \mathbb{Z} \quad (7.12)$$

with quotient groups

$$8\mathbb{Z} \cong \mathbb{Z}, \quad \mathbb{Z}_4 \quad \text{and} \quad \mathbb{Z}_2.$$



Otto Schreier (1901–1929)



Oberwolfach Photo Collection / Klaus Wohlfahrt

Hans Julius Zassenhaus (1912–1991) was born in Koblenz, the eldest son of Julius Zassenhaus, a historian and follower of the theologian and musician Albert Schweitzer (1875–1965). Hans entered the University of Hamburg in 1930, initially intending to specialise in theoretical physics, but he was inspired to pursue research in mathematics by Emil Artin (1898–1962), who became his doctoral supervisor.

At the age of 21, while a graduate student, he proved the celebrated Butterfly Lemma and used it to construct a neater proof of the Jordan–Hölder Theorem. After completing his doctoral thesis in 1934 (on a class of permutation groups now known as **Zassenhaus groups**) he taught for two years at the University of Rostock, before returning to Hamburg as Artin’s assistant, where in 1937 he wrote an influential textbook *Lehrbuch der Gruppentheorie* and completed his habilitation thesis on Lie rings in 1938.

He was strongly opposed to the Nazi regime, and his mother Margarete was involved in a resistance effort to hide endangered people. His refusal to join the Party resulted in the loss of his academic post in 1940, after which he worked on weather forecasting for the navy until the end of the war.

In 1949 he was appointed to a chair at McGill University in Montreal, during which time he worked on applications of computers to algebraic number theory. In 1959 he moved to Ohio State University, where he was based for the rest of his life. He retired in 1982 but continued writing papers and supervising research students until shortly before his death.

These series are clearly not isomorphic, but if we’re careful we can find refinements of each that are isomorphic. Each of these refinements should, at the very least, have quotient groups $\mathbb{Z}$, $\mathbb{Z}_2$, $\mathbb{Z}_3$ and $\mathbb{Z}_4$. We can refine (7.11) to get

$$\{0\} \triangleleft 24\mathbb{Z} \triangleleft 6\mathbb{Z} \triangleleft 3\mathbb{Z} \triangleleft \mathbb{Z}$$

which has quotient groups

$$24\mathbb{Z} \cong \mathbb{Z}, \quad \mathbb{Z}_4, \quad \mathbb{Z}_2 \quad \text{and} \quad \mathbb{Z}_3.$$

Similarly, we can refine (7.12) to get

$$\{0\} \triangleleft 24\mathbb{Z} \triangleleft 8\mathbb{Z} \triangleleft 2\mathbb{Z} \triangleleft \mathbb{Z},$$

which has quotient groups

$$24\mathbb{Z} \cong \mathbb{Z}, \quad \mathbb{Z}_3, \quad \mathbb{Z}_4 \quad \text{and} \quad \mathbb{Z}_2.$$

Both of these refinements are therefore isomorphic.

The next theorem, due to the Austrian mathematician Otto Schreier (1901–1929), says that we can always do this: any two subnormal (or normal) series for a group G have isomorphic refinements.

We will just state the theorem for the moment: the proof is somewhat involved and will require a bit of preparation.

Theorem 7.27 (Schreier’s Refinement Theorem) *Let G be a group, and let $H = \{H_0, \dots, H_m\}$ and $K = \{K_0, \dots, K_n\}$ be two subnormal (or normal) series for G . Then there exist refinements $S = \{S_0, \dots, S_l\}$ of H and $T = \{T_0, \dots, T_l\}$ of K such that $S \cong T$.*

To prove this, we first need a technical result due to the German mathematician Hans Zassenhaus (1912–1991), and in order to prove that we need a lemma attributed to the German mathematician Richard Dedekind (1831–1916).

Lemma 7.28 (Dedekind’s Modular Law) *Let G be a group, and suppose that X , Y and Z are subgroups of G , with $Z \subseteq X$. Then*

$$X \cap (YZ) = (X \cap Y)Z.$$

Proof Let $x \in X \cap (YZ)$. Then $x = yz$ for some $y \in Y$ and $z \in Z \subseteq X$, so $y = xz^{-1} \in X$, and hence $y \in X \cap Y$. Therefore $x \in (X \cap Y)Z$, and so $X \cap (YZ) \subseteq (X \cap Y)Z$.

Conversely, suppose that $x \in (X \cap Y)Z$. Then $x = ab$ for some $a \in X \cap Y$ and $b \in Z \subseteq X$, which means that $x \in X$. Also, $X \cap Y \subseteq Y$, so $(X \cap Y)Z \subseteq YZ$, and hence $x \in YZ$. Thus $x \in X \cap (YZ)$, so $(X \cap Y)Z \subseteq X \cap (YZ)$, and therefore $X \cap (YZ) = (X \cap Y)Z$. $\square$

The next result we need is Zassenhaus’ Lemma, sometimes called the “Butterfly Lemma” due to the shape of the subgroup lattice formed by

the various subgroups involved in the proof (see Figure 7.2). Zassenhaus formulated and proved his Lemma at the age of 21, while a graduate student at the University of Hamburg, in order to provide a neater proof of Schreier's Refinement Theorem and the Jordan–Hölder Theorem.

Theorem 7.29 (Zassenhaus' Lemma) *Let G be a group, and let H, K, A and B be subgroups of G such that $A \trianglelefteq H$ and $B \trianglelefteq K$. Then*

- (i) $A(H \cap B) \trianglelefteq A(H \cap K)$,
- (ii) $B(A \cap K) \trianglelefteq B(H \cap K)$, and
- (iii) $\frac{A(H \cap K)}{A(H \cap B)} \cong \frac{H \cap K}{(A \cap K)(H \cap B)} \cong \frac{B(H \cap K)}{B(A \cap K)}$.

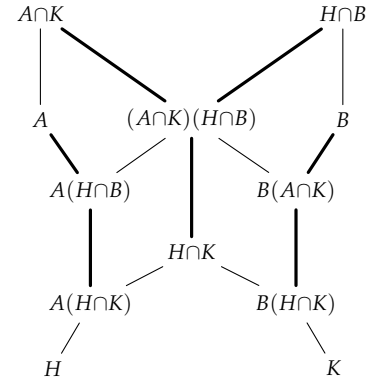


Figure 7.2: Subgroup lattice for Zassenhaus' Lemma

Figure 7.2 shows the subgroup lattice formed by the subgroups involved in the Lemma. For aesthetic reasons this lattice is drawn upside down relative to our usual convention: larger subgroups are towards the bottom of the diagram, while smaller subgroups are towards the top. Bold lines indicate that the upper subgroup is normal in the lower one. In general, a subgroup at the apex of two upward lines is the intersection of the subgroups at the other ends. Similarly, two downward lines meet at a subgroup that is the product of the subgroups at their other ends.

Proof To prove part (i), we must show first that $A(H \cap B)$ is a subgroup of $A(H \cap K)$, and then that it is closed under conjugation by elements of $A(H \cap K)$. By Proposition 2.11, $A(H \cap B) \leq A(H \cap K)$ if and only if $A(H \cap B) = (H \cap B)A$. But we know that $A \trianglelefteq H$, so $Ah = hA$ for any element $h \in H$, and hence this also follows for any $h \in H \cap B \subseteq H$. Therefore $A(H \cap B) = (H \cap B)A$ and thus $A(H \cap B) \leq A(H \cap K)$.

Now we must show that $A(H \cap B)$ is normal in $A(H \cap K)$, which we do by showing that it is closed under conjugation by any element of $A(H \cap K)$. Let $x = ab$ and $y = ck$ where $a, c \in A$, $b \in H \cap B$ and $k \in H \cap K$, so that $x \in A(H \cap B)$ and $y \in A(H \cap K)$. We want to show that $xyx^{-1} \in A(H \cap B)$.

First, observe that $yay^{-1} \in A$ since $y \in A(H \cap K) \subseteq AH \subseteq H$, and $A \trianglelefteq H$. Next, $kbk^{-1} \in H \cap B$ since $k \in H \cap K$ and $H \cap B \trianglelefteq H \cap K$. Hence $yby^{-1} = (ck)b(ck)^{-1} = ckbk^{-1}c^{-1} \in A(H \cap B)A$, and $A(H \cap B)A = A(H \cap B)$ since $A \trianglelefteq H$. Finally, $xyx^{-1} = y(ab)y^{-1} = (yay^{-1})(yby^{-1}) \in A(H \cap B)$, and therefore $A(H \cap B) \trianglelefteq A(H \cap K)$ as claimed.

Part (ii) follows by a very similar argument.

To prove part (iii), we use Dedekind's Modular Law¹⁴ and the Second Isomorphism Theorem.¹⁵ Recall that the latter says that if a group G contains a subgroup M and a normal subgroup N , then $MN/N \cong$

¹⁴ Lemma 7.28, page 248.

¹⁵ Theorem 4.62, page 126.

$M/(M \cap N)$. Now set $G = A(H \cap K)$, and let $M = H \cap K$ and $N = A(H \cap B)$, which is normal in G by part (i). The Second Isomorphism Theorem says that

$$\frac{(H \cap K)A(H \cap B)}{A(H \cap B)} \cong \frac{H \cap K}{(H \cap K) \cap A(H \cap B)}. \quad (7.13)$$

According to Figure 7.2, $(H \cap K)A(H \cap B)$ should be equal to $A(H \cap K)$. To see this, first we observe that A is normal in H , and so $hA = Ah$ for any element $h \in H$. Consequently, since $H \cap K \subseteq H$, we also have $hA = Ah$ for any $h \in H \cap K$, and hence $(H \cap K)A = A(H \cap K)$, which in turn means that $(H \cap K)A(H \cap B) = A(H \cap K)(H \cap B)$. Next, observe that $(H \cap K)(H \cap B)$ consists of all products of the form xy where $x \in H \cap K$ and $y \in H \cap B$. But since $H \cap K \subseteq H \cap B$, it follows that $(H \cap K)(H \cap B) = H \cap K$, and hence $(H \cap K)A(H \cap B) = A(H \cap K)$ as claimed.

Now we apply Dedekind's Modular Law to $(H \cap K) \cap A(H \cap B)$. Setting $X = H \cap K$, $Y = A$ and $Z = H \cap B$, by the identity $X \cap (YZ) = (X \cap Y)Z$ we get

$$(H \cap K) \cap A(H \cap B) = (H \cap K \cap A)(H \cap B) = (A \cap K)(H \cap B),$$

and hence (7.13) becomes

$$\frac{A(H \cap K)}{A(H \cap B)} \cong \frac{H \cap K}{(A \cap K)(H \cap B)}. \quad (7.14)$$

By a very similar argument we find also that

$$\frac{B(H \cap K)}{B(A \cap K)} \cong \frac{H \cap K}{(A \cap K)(H \cap B)}. \quad (7.15)$$

Putting (7.14) and (7.15) together yields the required isomorphism. $\square$

We are now ready to prove Schreier's Refinement Theorem. The key idea of this proof is to insert between each group H_i and H_{i+1} in the series H a chain of groups of the form $H_i(H_{i+1} \cap K_j)$, and then to perform the analogous construction with the series K . We then apply Zassenhaus' Lemma to show the existence of a bijective correspondence between the relevant quotient groups.

Proof of Schreier's Refinement Theorem We first prove the theorem for subnormal series. Suppose that H and K are two subnormal series for a group G , with

$$\begin{aligned} \{e\} &= H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_m = G \\ \text{and} \quad \{e\} &= K_0 \triangleleft K_1 \triangleleft \cdots \triangleleft K_n = G. \end{aligned}$$

Now consider the chain

$$H_i(H_{i+1} \cap K_0) \leq H_i(H_{i+1} \cap K_1) \leq \cdots \leq H_i(H_{i+1} \cap K_n).$$

This consists of $(n+1)$ not necessarily distinct subgroups of G . Furthermore,

$$H_i(H_{i+1} \cap K_0) = H_i(H_{i+1} \cap \{e\}) = H_i\{e\} = H_i$$

and $H_i(H_{i+1} \cap K_n) = H_i(H_{i+1} \cap G) = H_i H_{i+1} = H_{i+1}.$

So, between each pair of subgroups H_i and H_{i+1} in the first series H , we insert the chain

$$H_i(H_{i+1} \cap K_1) \leq H_i(H_{i+1} \cap K_2) \leq \cdots \leq H_i(H_{i+1} \cap K_{n-1}).$$

This yields a chain of $mn+1$ subgroups of G , not all of which need necessarily be distinct.

Similarly, between each pair of subgroups K_j and K_{j+1} in the second series K , we insert the chain

$$K_j(K_{j+1} \cap H_1) \leq K_j(K_{j+1} \cap H_2) \leq \cdots \leq K_j(K_{j+1} \cap H_{m-1}).$$

This also results in a sequence of $mn+1$ subgroups of G , which again need not all be distinct.

By parts (i) and (ii) of Zassenhaus' Lemma, each subgroup in each of these chains is normal in the next subgroup along. That is,

$$H_i(H_{i+1} \cap K_j) \trianglelefteq H_i(H_{i+1} \cap K_{j+1})$$

$$\text{and } K_j(K_{j+1} \cap H_i) \trianglelefteq K_j(K_{j+1} \cap H_{i+1})$$

for all $0 \leq i < m$ and $0 \leq j < n$, since each $H_i \triangleleft H_{i+1}$ and $K_j \triangleleft K_{j+1}$.

Also, by part (iii) of Zassenhaus' Lemma, we have isomorphisms

$$\frac{H_i(H_{i+1} \cap K_{j+1})}{H_i(H_{i+1} \cap K_j)} \cong \frac{K_j(K_{j+1} \cap H_{i+1})}{K_j(K_{j+1} \cap H_i)}. \quad (7.16)$$

These isomorphisms give a bijective correspondence between the $mn+1$ groups in the chain constructed from H , and the $mn+1$ groups in the chain formed from K .

But in Definition 7.23 we stipulated that in a series all inclusions must be strict, so we now have to discard all the repeated subgroups from these two chains. The quotient of two repeated subgroups is clearly trivial, so each trivial quotient in the chain obtained from H corresponds with a trivial quotient in the chain we constructed from K , and vice versa. So discarding repeated subgroups is the same as discarding trivial quotients, and there must be the same number in each chain.

By doing this, the chain constructed from H yields a series $S = \{S_0, \dots, S_l\}$ that is a refinement of H , and the chain obtained from K becomes a series $T = \{T_0, \dots, T_l\}$ that is a refinement of K . Furthermore, S and T are isomorphic.

All that remains is to confirm this works for normal series as well. But if H and K are normal series, then each group $H_i(H_{i+1} \cap K_j)$ and

$K_j(K_{j+1} \cap H_i)$ is normal in G , and hence the refinements constructed by this method will also be normal. $\square$

To see an example of this in practice, let's consider two different, non-equivalent subnormal series of the dihedral group D_4 .

Example 7.30 Let

$$H_0 = \{e\}, \quad H_1 = \{e, r, r^2, r^3\}, \quad H_2 = D_4$$

and

$$K_0 = \{e\}, \quad K_1 = \{e, m_1\}, \quad K_2 = \{e, r^2, m_1, m_3\}, \quad K_3 = D_4$$

be two subnormal series for D_4 . We refine H by inserting the groups

$$H_0(H_1 \cap K_1) = H_0 \quad \text{and} \quad H_0(H_1 \cap K_2) = \{e, r^2\}$$

between H_0 and H_1 , and the groups

$$H_1(H_2 \cap K_1) = D_4 \quad \text{and} \quad H_1(H_2 \cap K_2) = D_4$$

between H_1 and H_2 . This gives the chain

$$H_0 \leq H_0 \leq \{e, r^2\} \leq H_1 \leq D_4 \leq D_4 \leq D_4$$

which, after discarding repeated groups, yields the series

$$\{e\} \triangleleft \{e, r^2\} \triangleleft \{e, r, r^2, r^3\} \triangleleft D_4.$$

This series has quotients $\mathbb{Z}_2, \mathbb{Z}_2$ and $\mathbb{Z}_2$.

Similarly, we refine K by inserting the group

$$K_0(K_1 \cap H_1) = K_0$$

between the groups K_0 and K_1 , the group

$$K_1(K_2 \cap H_1) = K_2$$

between the groups K_1 and K_2 , and the group

$$K_2(K_3 \cap H_1) = D_4$$

between the groups K_2 and K_3 . This gives the chain

$$K_0 \leq K_0 \leq K_1 \leq K_2 \leq K_2 \leq D_4 \leq D_4,$$

which yields the series

$$\{e\} \triangleleft \{e, m_1\} \triangleleft \{e, r^2, m_1, m_3\} \triangleleft D_4.$$

This series is actually the same as K itself, which indicates that K can't be refined any further than it is: there is no room for any more normal subgroups between any of the existing groups. Its quotients are $\mathbb{Z}_2, \mathbb{Z}_2$ and $\mathbb{Z}_2$, which are the same as those in the refined series obtained from H , and hence the two refined series are isomorphic, as predicted by Schreier's Refinement Theorem.

In this example, the subnormal series K , that is,

$$\{e\} \triangleleft \{e, m_1\} \triangleleft \{e, r^2, m_1, m_3\} \triangleleft D_4$$

was already as long as it could be: there is no room to fit any more normal subgroups in between any of the others on the list. Each group is in some sense a largest possible proper normal subgroup of the group containing it.

Definition 7.31 Let G be a group, and suppose that $H < G$ is a proper subgroup of G that isn't contained in any other proper subgroup of G . That is, there exists no other subgroup $K < G$ such that $H < K$. Then we say H is a **maximal subgroup** of G .

If $N \triangleleft G$ is a normal subgroup of G that isn't contained in any other proper normal subgroup of G , then we call N a **maximal normal subgroup** of G .

Maximal normal subgroups have an important property:

Proposition 7.32 Let G be a group and $N \triangleleft G$ be a proper normal subgroup of G . Then G/N is simple exactly when N is a maximal normal subgroup of G .

Proof Let N be a maximal normal subgroup of G . The quotient homomorphism $q: G \rightarrow G/N$ maps every element $g \in G$ to the coset gN in G/N . By Proposition 4.37, the preimage $q^{-1}(H)$ of any proper normal subgroup of G/N would be a proper normal subgroup of G such that $N \triangleleft q^{-1}(H)$. But we said that N was maximal normal in G , so this can't happen and therefore G/N must be simple.

Now suppose that G/N is simple, and that there exists some proper normal subgroup $M \triangleleft G$ with $N \triangleleft M$. Then $q(M) \triangleleft G/N$. But G/N is simple, so no such proper normal subgroup $q(M)$ can exist, and hence M can't exist either, so N is maximal normal in G . $\square$

The reason we can't fit another subgroup in between, for example, $\{e, m_1\}$ and $\{e, r^2, m_1, m_3\}$ is because the former is a maximal normal subgroup of the latter. We want to study series of this type further.

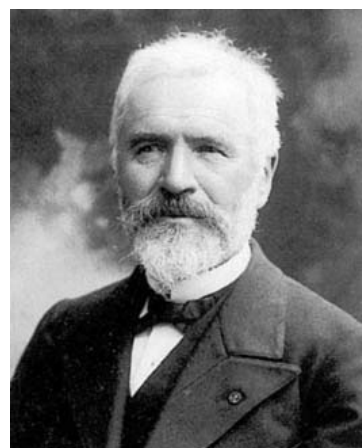
Definition 7.33 A subnormal series

$$\{e\} = H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_n = G$$

of a group G is a **composition series** if it cannot be refined any further. Equivalently, each group H_i is a maximal normal subgroup of H_{i+1} for $0 \leq i < n$. By Proposition 7.32, this is the same as requiring each quotient (or **composition factor**) H_{i+1}/H_i to be simple.

A normal series with this property is called a **principal** or **chief series**.

Not every group has a composition series or principal series. For



Wikimedia Commons

The French mathematician **Marie Ennemond Camille Jordan** (1838–1922) was born to a wealthy family in Lyon, and studied engineering and mathematics at the École Polytechnique, as his father had before him. He worked as an engineer for some years, doing mathematical research in his spare time, and obtaining his doctorate in 1861 for a two-part thesis entitled *Sur le nombre des valeurs des fonctions* and *Sur des périodes des fonctions inverses des intégrales des différentielles algébriques*.

He was appointed professor at the École Polytechnique in 1876, and also at the Collège de France in 1883, but continued to work as an engineer until 1885, during which period he wrote his influential textbook *Cours d'Analyse*, published between 1882 and 1887.

He made many contributions to several different areas of mathematics: as well as the Jordan–Hölder Theorem, he also formulated the Jordan Curve Theorem in topology and complex analysis, and the Jordan normal form in linear algebra. He also did much work on Galois theory and studied the Mathieu groups mentioned in Section 3.A. His 1870 monograph on permutation groups, *Traité des substitutions et des équations algébriques*, won the Prix Ponclet, a prestigious prize awarded by the Académie des Sciences.

Several other members of his family attained distinction in various fields: his uncle Pierre Puvis de Chavannes (1824–1898) was a renowned artist, a great uncle also named Camille Jordan (1771–1821) became an influential politician, and his cousin Alexis Jordan (1814–1897) was an eminent botanist.



Otto Ludwig Hölder (1859–1937)

example, consider the subnormal (and normal) series

$$\{0\} \triangleleft 120\mathbb{Z} \triangleleft 60\mathbb{Z} \triangleleft 12\mathbb{Z} \triangleleft 6\mathbb{Z} \triangleleft 2\mathbb{Z} \triangleleft \mathbb{Z}.$$

There is no room for any more subgroups between $120\mathbb{Z}$ and $\mathbb{Z}$: we can see this by considering the quotients $\mathbb{Z}_2, \mathbb{Z}_5, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_2$, which are all cyclic groups of prime order and hence simple. But we can certainly fit another group in between $\{0\}$ and $120\mathbb{Z}$: for example, $240\mathbb{Z}$, which yields another simple quotient $\mathbb{Z}_2$. And having done so, we can insert another group, say $720\mathbb{Z}$, before that, and so on.

The problem, though, is that whatever subgroup we insert here will have to be of the form $k\mathbb{Z}$ for some $k \in \mathbb{Z}$, and so the first quotient will always be isomorphic to $k\mathbb{Z}$, which is not simple. A little thought will hopefully convince you that this problem arises for any subnormal series of $\mathbb{Z}$, so we're forced to conclude that $\mathbb{Z}$ has no composition series (and hence it can't have a principal series either).

But those groups that do have composition series satisfy an important theorem originally proved by the French mathematician Camille Jordan (1838–1922) and later extended by the German mathematician Otto Hölder (1859–1937).

Theorem 7.34 (The Jordan–Hölder Theorem) *Any two composition (or principal) series for a group G are isomorphic.*

Proof Let

$$\begin{aligned} \{e\} &= H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_m = G \\ \text{and} \quad \{e\} &= K_0 \triangleleft K_1 \triangleleft \cdots \triangleleft K_n = G \end{aligned}$$

be two composition (or principal) series for G . Then by Schreier's Refinement Theorem both of these series have isomorphic refinements. But since all the composition factors are already simple, each group H_i is maximal normal in H_{i+1} and each group K_j is maximal normal in K_{j+1} for $0 \leq i < m$ and $0 \leq j < n$, and hence neither series can be refined any further. Therefore both series must already be isomorphic. $\square$

Also in Example 7.30 we found another composition series for D_4 , namely

$$\{e\} \triangleleft \{e, r^2\} \triangleleft \{e, r, r^2, r^3\} \triangleleft D_4.$$

This series contains the rotation subgroup $R_4 = \{e, r, r^2, r^4\}$, which is known to be normal in D_4 . In fact, if a group G has a composition series at all, then we can find one that contains any given normal subgroup $N \triangleleft G$:

Proposition 7.35 *Let G be a group that has a composition (or principal) series, and let $N \triangleleft G$ be a proper normal subgroup of G . Then there exists some composition (or principal) series for G that contains N .*

Proof Suppose that

$$\{e\} = H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_n = G$$

is a composition (or principal) series for G . The series

$$\{e\} \triangleleft N \triangleleft G$$

is both subnormal and normal. By Schreier's Refinement Theorem there is a refinement of this series that is isomorphic to the given composition (or principal) series, and hence itself a composition (or principal) series for G . This refinement will necessarily contain N . $\square$

7.3 Soluble and nilpotent groups

PARTICULARLY IMPORTANT is the case where the composition factors are abelian. This will become especially relevant in Chapter 11 when we study the solubility by radicals of polynomial equations. The only abelian finite simple groups are the cyclic groups $\mathbb{Z}_p$ where p is prime, so this condition is equivalent to requiring the composition factors to be finite cyclic groups of prime order.

Definition 7.36 A group G is **soluble** or **solvable** if it has a composition series whose composition factors are all abelian, or equivalently are cyclic groups $\mathbb{Z}_p$ of prime order.

From earlier discussion we know that D_4 is soluble, since it has a composition series

$$\{e\} \triangleleft \{e, m_1\} \triangleleft \{e, r^2, m_1, m_3\} \triangleleft D_4$$

whose composition factors are all isomorphic to $\mathbb{Z}_2$. The group $\mathbb{Z}_{30}$ is soluble since it has a composition series

$$\{0\} \triangleleft \langle 6 \rangle \triangleleft \langle 2 \rangle \triangleleft \mathbb{Z}_{30}$$

with composition factors $\mathbb{Z}_5$, $\mathbb{Z}_3$ and $\mathbb{Z}_2$, all of which are simple and abelian. The symmetric group S_5 has a composition series

$$\{1\} \triangleleft A_5 \triangleleft S_5,$$

which has composition factors A_5 and $\mathbb{Z}_2$. The alternating group A_5 is simple but not abelian, so S_5 is not soluble. An important consequence of this is the fact, originally discovered independently in the early

The world is devoted to physical science, because it believes these discoveries will increase its capacity of luxury and self-indulgence. But the pursuit of science only leads to the insoluble. When we arrive at that barren term, the Divine voice summons man, as it summoned Samuel;

— Benjamin Disraeli, 1st Earl of Beaconsfield (1804–1881), *Lothair* (1870) 70

19th century by Évariste Galois (1811–1832) and Niels Henrik Abel (1802–1829), that the general quintic equation

$$ax^5 + bx^4 + cx^3 + dx^2 + ex + f = 0$$

cannot be solved by radicals.

In fact, apart from the first few cases, no finite symmetric group S_n is soluble:

Proposition 7.37 *The symmetric group S_n is not soluble for $n \geq 5$.*

Proof By Proposition 3.43, the alternating group A_n is simple for $n \geq 5$. And by the discussion in Example 3.33, A_n is normal in S_n . We thus have a subnormal series

$$\{1\} \triangleleft A_n \triangleleft S_n$$

with factors A_n and $\mathbb{Z}_2$, both of which are simple, and so this is a composition series for S_n . By the Jordan–Hölder Theorem, any other composition series for S_n will also have composition factors isomorphic to A_n and $\mathbb{Z}_2$, and since A_n is nonabelian, S_n isn't soluble. $\square$

However, lots of important classes of groups are soluble. The next few results show that subgroups, quotients and direct products of soluble groups are also soluble.

Proposition 7.38 *If a group G is soluble, then any subgroup $H \leq G$ is also soluble.*

Proof Suppose that

$$\{e\} = G_0 \triangleleft G_1 \triangleleft \cdots \triangleleft G_n = G$$

is a composition series for G . Now form another chain of subgroups

$$\{e\} = H_0 < H_1 < \cdots < H_n = H \quad (7.17)$$

where $H_i = (H \cap G_i)$ for $0 \leq i \leq n$. By Lemma 4.61 we know that $H_i \triangleleft H_{i+1}$ for $0 \leq i < n$ since

$$H_i = H \cap G_i = (H \cap G_i) \cap G_{i+1} \triangleleft H \cap G_{i+1} = H_{i+1},$$

so (7.17) is a subnormal series for H .

Now set $A = \{e\}$, $K = G_{i+1}$ and $B = G_i$ and apply Zassenhaus' Lemma to see that

$$\begin{aligned} \frac{H_{i+1}}{H_i} &= \frac{H \cap G_{i+1}}{H \cap G_i} = \frac{H \cap K}{(A \cap K)(H \cap B)} \cong \\ &\quad \frac{B(H \cap K)}{B(A \cap K)} = \frac{G_i(H \cap G_{i+1})}{G_i} < \frac{G_{i+1}}{G_i}. \end{aligned}$$

Since G_{i+1}/G_i is abelian, its subgroup H_{i+1}/H_i must be as well, and therefore H is soluble. $\square$

Proposition 7.39 *If a group G is soluble, and $N \triangleleft G$ is normal in G , then the quotient G/N is also soluble.*

Proof By Proposition 7.35 there is a composition series

$$\{e\} = H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_k = N \triangleleft \cdots \triangleleft H_n = G$$

for G that contains N . Taking the quotient of each of the groups $H_k, \dots, H_n$ by N we obtain a new series

$$\{e\} = N/N = H_k/N < H_{k+1}/N < \cdots < H_n/N = G/N.$$

By the Third Isomorphism Theorem¹⁶ this is a subnormal series, and $(H_{i+1}/N)/(H_i/N) \cong H_{i+1}/H_i$ for $k \leq i < n$. Each of these quotients must therefore be abelian, and hence this new series is a composition series for G/N with abelian composition factors. Therefore G/N is soluble. $\square$

¹⁶ Theorem 4.65, page 128.

Proposition 7.40 *Let $N \triangleleft G$ be a normal subgroup of some group G . If N and G/N are both soluble, then so is G .*

Proof Since N and G/N are soluble, there exist composition series

$$\{e\} = H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_m = N$$

and

$$\{e\} = K_0 \triangleleft K_1 \triangleleft \cdots \triangleleft K_n = G/N.$$

In the latter series, each group K_i can be written as G_i/N for some subgroup G_i , and N corresponds to the identity element in the quotient group G/N , so this can be written as

$$N \triangleleft G_1/N \triangleleft \cdots \triangleleft G_n/N = G/N.$$

Each of the composition factors of both of these series are cyclic groups of prime order, and by the Third Isomorphism Theorem we have $(G_{i+1}/N)/(G_i/N) \cong G_{i+1}/G_i$, which must also be of prime order. We can therefore construct a composition series

$$\{e\} = H_0 \triangleleft H_1 \triangleleft \cdots \triangleleft H_m = N \triangleleft G_1 \triangleleft \cdots \triangleleft G_n = G$$

for G , all of whose composition factors are cyclic groups of prime order, and hence abelian. Thus G is soluble. $\square$

An important corollary of this last proposition is that direct products of soluble groups are themselves soluble:

Corollary 7.41 *If H and K are soluble groups, then so is $H \times K$.*

Proof Let $G = H \times K$. Then $H \triangleleft G$ and $G/H \cong K$, so by Proposition 7.40 it follows that G must also be soluble. $\square$

We can use Proposition 7.40 to show first that finite abelian groups are soluble, and then that finite p -groups are soluble too.

Proposition 7.42 *Let G be a finite abelian group. Then G is soluble.*

Proof Suppose that $|G| = p$ for some prime integer p . Then $G \cong \mathbb{Z}_p$, which is soluble, since there is a composition series

$$\{0\} \triangleleft \mathbb{Z}_p$$

with composition factor isomorphic to $\mathbb{Z}_p$ itself.

Otherwise, we proceed by induction on $n = |G|$, which we assume to be composite. Suppose that all abelian groups of order less than n have already been shown to be soluble. If p is a prime factor of $|G|$, then by Cauchy's Theorem¹⁷ G must have a cyclic subgroup $H \cong \mathbb{Z}_p$. This subgroup H is normal in G , and its quotient G/H is abelian with order $|G/H| = n/p < n$. Therefore both H and G/H are soluble, and by Proposition 7.40, so is G . $\square$

¹⁷ Theorem 2.37, page 57.

Proposition 7.43 *Let G be a finite p -group of order $|G| = p^k$ for some prime p and positive integer k . Then G is soluble.*

Proof If $k = 1$ then $|G| = p$ and hence $G \cong \mathbb{Z}_p$, which is soluble. We now proceed by induction on k . By Proposition 7.19, the centre $Z(G)$ is nontrivial, and by Proposition 3.17, $Z(G) \triangleleft G$. Furthermore, the centre $Z(G)$ is abelian, and hence soluble by Proposition 7.42. Also $G/Z(G)$ is a p -group of order p^m , with $m < k$, and is hence soluble by the inductive hypothesis. Therefore, by Proposition 7.40 it follows that G must also be soluble. $\square$

There are two other important results that we will now state without proof. The first of these is another important result named after (and in this case originally proved by) the British group theorist William Burnside (1852–1927).

Theorem 7.44 (Burnside's Theorem) *Any group G of order $p^m q^n$, where p and q are both prime and m and n are non-negative integers, is soluble.*

The usual proof of this theorem involves some sophisticated techniques from representation theory. Also worth mentioning is a celebrated result due to the American mathematicians Walter Feit (1930–2004) and John Griggs Thompson. The published proof of this theorem is rather complicated and filled an entire 255-page issue of the *Pacific Journal of Mathematics*¹⁸ so we will omit it.

¹⁸ W Feit and JG Thompson, *Solvability of groups of odd order*, *Pacific Journal of Mathematics* 13.3 (1963) 775–1029.

Theorem 7.45 (Feit–Thompson Theorem) *Any group of odd order is soluble.*

One important normal subgroup we met in Chapter 3 was the **commutator subgroup** or **derived subgroup** $[G, G]$ of a group G .¹⁹ This gives some measure of how nonabelian a group is: $[G, G]$ is trivial exactly

¹⁹ Definition 3.19, page 78.

when G is abelian, and the larger $[G, G]$, the more G fails to be abelian. For any abelian group G , then, we have a very short subnormal series

$$[G, G] = \{e\} \triangleleft G.$$

But what happens if G isn't abelian? If we try this for S_3 , the smallest nonabelian group, we find that $[S_3, S_3] = A_3 = \langle (123) \rangle \cong \mathbb{Z}_3$, which yields the subnormal series

$$\{1\} \triangleleft [S_3, S_3] = A_3 \triangleleft S_3.$$

Since A_3 is abelian, its commutator subgroup is trivial, so what we have here is a series of subgroups of S_3 , in which each group is the commutator subgroup of the next group on the list. More generally, we can get a descending series of groups in which each is the commutator subgroup of its predecessor:

Definition 7.46 For a group G , recursively define subgroups

$$G^{(0)} = G, \quad G^{(1)} = [G^{(0)}, G^{(0)}], \quad G^{(2)} = [G^{(1)}, G^{(1)}], \quad \dots$$

such that $G^{(i+1)} = [G^{(i)}, G^{(i)}]$ for $i \geq 0$. This series

$$\dots \triangleleft G^{(2)} \triangleleft G^{(1)} \triangleleft G^{(0)} = G$$

is called the **derived series** of G .

The derived series for S_4 is

$$\{1\} \triangleleft \langle (12)(34), (13)(24) \rangle \triangleleft A_4 \triangleleft S_4. \quad (7.18)$$

But if we try this for the symmetric group S_5 , we find that $[S_5, S_5] = A_5$,²⁰ but then we run into a problem (or, synonymously, some interesting behaviour) because A_5 is simple.²¹ By Proposition 3.21, commutator subgroups are normal, so $[A_5, A_5]$ must be either trivial or A_5 itself. But A_5 isn't abelian, so its commutator subgroup can't be trivial, which means that $[A_5, A_5] = A_5$. The derived series for S_5 , then, doesn't terminate at the trivial subgroup:

$$A_5 \triangleleft S_5.$$

The other property S_5 has that is relevant to the current discussion is that it isn't soluble. This isn't a coincidence:

Proposition 7.47 *A group G is soluble if and only if its derived series terminates at the trivial subgroup. That is, if $G^{(n)} = \{e\}$ for some $n \geq 0$.*

Proof Suppose that

$$\{e\} = G^{(n)} \triangleleft G^{(n-1)} \triangleleft \dots \triangleleft G^{(1)} \triangleleft G^{(0)} = G$$

is the derived series for G . Then each quotient $G^{(i)} / G^{(i+1)}$ is abelian. In particular, $G^{(n-1)} / G^{(n)} = G^{(n-1)}$ is abelian and hence soluble. Next, $G^{(n-2)} / G^{(n-1)}$ is abelian and hence soluble, so by Proposition 7.40, $G^{(n-2)}$ must also be soluble. We proceed by reverse induction on n :

²⁰ Proposition 3.20, page 78.

²¹ Proposition 3.43, page 92.

having shown that $G^{(k)}$ is soluble, and since $G^{(k-1)}/G^{(k)}$ is abelian and therefore soluble, Proposition 7.40 implies that $G^{(k-1)}$ must also be soluble. Therefore, by induction, $G = G^{(0)}$ is soluble.

Conversely, suppose G is soluble. Then there is a composition series

$$\{e\} = G_m \triangleleft G_{m-1} \triangleleft \cdots \triangleleft G_0 = G$$

for G , with each composition factor G_i/G_{i+1} an abelian simple group, for $0 \leq i \leq m$. This composition series will be at least as long as the derived series, and we claim that $G^{(i)} \leq G_i$ for $0 \leq i \leq m$.

Meanwhile, the derived groups eventually stabilise, in the sense that there exists a positive integer n such that $G^{(i+1)} = G^{(i)}$ for all $i \geq n$.

We proceed by induction on i : for $i = 0$ we have $G^{(0)} = G = G_0$, and since G_0/G_1 is abelian, $G^{(1)} = [G, G] \leq G_1$ by Proposition 3.31. Suppose that $G^{(i)} \leq G_i$ for some i . The quotient G_i/G_{i+1} is abelian, so $[G_i, G_i] \leq G_{i+1}$ by Proposition 3.31, and hence

$$G^{(i+1)} = [G^{(i)}, G^{(i)}] \leq [G_i, G_i] \leq G_{i+1}$$

as claimed. Therefore $G^{(n)} \leq G_n = \{e\}$ and thus the derived series terminates at the trivial subgroup. $\square$

This gives another way of testing for solubility: construct the derived series and see if it eventually terminates at the trivial subgroup. Moreover, the derived series of a group is the shortest possible subnormal series with abelian quotients, and its length thus gives us some potentially useful information about the complexity of the group:

Definition 7.48 Let G be a soluble group. Then the length of the derived series for G is the **derived length** of the group.

The trivial group $\{e\}$ is the only group with derived length 0. The groups of derived length 1 are exactly the nontrivial abelian groups: if A is abelian and soluble, then $A^{(0)} = A$, while $A^{(1)} = [A, A] = \{e\}$; conversely if $A^{(1)} = [A, A] = \{e\}$ then this means that every element of A commutes with every other element of A , and therefore A is abelian. The groups of derived length 2 are precisely those with nontrivial abelian commutator groups $G^{(1)} = [G, G]$; such groups are called **metabelian**, and we will briefly return to them later.

Another important normal subgroup we met in Chapter 3 is the centre $Z(G)$ of a group G . We can use it to construct normal or subnormal series but in a slightly more complicated way than we did with the commutator subgroup $[G, G]$. The obvious method, where each group is the centre of its predecessor, doesn't work, or at least not in a very interesting way: $Z(G)$ is abelian, so its centre is the same; that is, $Z(Z(G)) = Z(G)$. Hence any such series will stabilise after at most one step.

So we need a different approach, and the key is to consider the centre of quotient groups instead. More precisely:

Definition 7.49 A normal series

$$\{e\} = G_0 < G_1 < \cdots < G_n = G$$

is **central** if $G_{i+1}/G_i \leq Z(G/G_i)$ for $0 \leq i < n$.

That is, each quotient G_{i+1}/G_i of the series is contained in the centre of the corresponding quotient G/G_i of the full group G . For example:

Example 7.50 Suppose that

$$\{e\} \triangleleft \{e, r^2\} \triangleleft \{e, r, r^2, r^3\} \triangleleft D_4 \quad (7.19)$$

is a normal series for D_4 , with

$$G_0 = \{e\}, \quad G_1 = \{e, r^2\}, \quad G_2 = \{e, r, r^2, r^3\}, \quad G = G_3 = D_4.$$

The quotient G_3/G_2 consists of two cosets

$$\begin{aligned} eG_2 = rG_2 = r^2G_2 = r^3G_2 &= \{e, r, r^2, r^3\}, \\ m_1G_2 = m_2G_2 = m_3G_2 = m_4G_2 &= \{m_1, m_2, m_3, m_4\}. \end{aligned}$$

This is the same as the quotient G/G_2 , whose centre $Z(G/G_2)$ consists of all cosets that commute with everything else. But since $G/G_2 \cong \mathbb{Z}_2$ is abelian, we have

$$G_3/G_2 = Z(G/G_2).$$

The quotient G_2/G_1 consists of the cosets

$$eG_1 = r^2G_1 = \{e, r^2\}, \quad rG_1 = r^3G_1 = \{r, r^3\}.$$

Meanwhile, the quotient G/G_1 consists of the cosets

$$\begin{aligned} eG_1 = r^2G_1 &= \{e, r^2\}, & rG_1 = r^3G_1 &= \{r, r^3\}, \\ m_1G_1 = m_3G_1 &= \{m_1, m_3\}, & m_2G_1 = m_4G_1 &= \{m_2, m_4\}. \end{aligned}$$

This group is of order 4 and hence abelian, so $Z(G/G_1) = G/G_1$, and we can also see that $G_2/G_1 \leq Z(G/G_1)$.

Finally, G_1/G_0 consists of the cosets

$$eG_0 = \{e\}, \quad r^2G_0 = \{r^2\},$$

while G/G_0 consists of the cosets

$$\begin{aligned} eG_0 &= \{e\}, & rG_0 &= \{r\}, & r^2G_0 &= \{r^2\}, & r^3G_0 &= \{r^3\}, \\ m_1G_0 &= \{m_1\}, & m_2G_0 &= \{m_2\}, & m_3G_0 &= \{m_3\}, & m_4G_0 &= \{m_4\}. \end{aligned}$$

The centre of this quotient is $\{eG_0, r^2G_0\}$, and thus again we have $G_1/G_0 \leq Z(G/G_0)$. So (7.19) is a central series.

This construction only works with a normal series, since we need each quotient G/G_i to be a valid quotient group. It will always be

the case that G_{i+1}/G_i is a subgroup of G/G_i , but the requirement that $G_{i+1}/G_i \leq Z(G/G_i)$ is stronger, and will not always hold. A central series will always have abelian quotients G_{i+1}/G_i but not every series with abelian quotients need be central.

There are two ways to form a central series from a group G : start at $\{e\}$ and work upwards to construct an ascending series, or start at G itself and recursively form a descending series. These methods need not result in the same series; also the ascending series need not terminate at G , and the descending series might not reach $\{e\}$.

We'll try both of these approaches in turn. First we'll build an ascending series starting at $G_0 = \{e\}$. We want the second group G_1 to have the property that $G_1/G_0 \leq Z(G/G_0)$, and one way of ensuring this is to choose G_1 such that $G_1/G_0 = Z(G/G_0)$. That is, we want G_1 to be the subgroup of G whose quotient by G_0 is exactly the centre of G/G_0 , and we can achieve this by setting $G_1 = Z(G)$.

Next we want G_2 to be a normal subgroup of G such that $G_2/G_1 \leq Z(G/G_1)$ and again we'll choose G_2 so that $G_2/G_1 = Z(G/G_1)$. To do this, we need to set G_2 to be the subgroup of G whose quotient by G_1 is exactly the centre of the quotient G/G_1 . Continuing this process we obtain an ascending central series of groups:

Definition 7.51 For a group G , form an ascending normal series

$$G_0 = \{e\} \triangleleft G_1 \triangleleft \cdots \triangleleft G_n$$

such that $G_1 = Z(G)$ and G_i is the **higher centre** $Z^i(G)$ of G defined such that $Z^0(G) = \{e\}$, and

$$Z^i(G)/Z^{i-1}(G) = Z(G/Z^{i-1}(G))$$

for $0 < i \leq n$. This series is the **ascending central series** or **upper central series** of G .

As an example, we'll do this for the dihedral group D_4 :

Example 7.52 Let $G = D_4$. Then $G_0 = Z^0(G) = \{e\}$ and $G_1 = Z^1(G) = Z(G) = \{e, r^2\}$. Next we want $G_2 = Z^2(G)$ to be the subgroup of G such that $G_2/G_1 = Z(G/G_1)$. Now G/G_1 consists of the cosets

$$\begin{aligned} eG_1 = r^2G_1 &= \{e, r^2\}, & rG_1 &= r^3G_1 = \{r, r^3\}, \\ m_1G_1 = m_3G_1 &= \{m_1, m_3\}, & m_2G_1 &= m_4G_1 = \{m_2, m_4\}. \end{aligned}$$

This quotient group is of order 4 and hence abelian, so $Z(G/G_1) = G/G_1$, and hence $G_2 = Z^2(G)$ must be G itself. Therefore the upper central series of D_4 is

$$\{e\} \triangleleft \{e, r^2\} \triangleleft D_4.$$

As remarked earlier, the upper central series need not terminate at G . If we try this construction with the symmetric group $G = S_3$, for example, we find that $G_0 = Z^0(G) = \{1\}$, and $G_1 = Z^1(G) = Z(G) = \{1\}$ again. So all the higher centres of S_3 are trivial, and thus the upper central series consists of just the trivial subgroup $\{1\}$.

The upper central series for the dihedral group D_6 is a little more interesting, but also doesn't reach D_6 :

$$\{e\} \triangleleft \{e, r^3\}.$$

We'll explore this idea further in a little while, but now it's time to try the other approach and construct a descending normal series by starting at G and working downwards.

So, we set $G_0 = G$ to begin with. Next we want G_1 to have the property $G_0/G_1 \leq Z(G/G_1)$ and as before the obvious way of achieving this is to choose G_1 such that $G_0/G_1 = Z(G/G_1)$. What this means is that G_1 has to be selected so that G/G_1 is abelian, and the obvious way of doing this is to set $G_1 = [G, G]$. Any other choice of G_1 for which G/G_1 is abelian will contain $[G, G]$, by Proposition 3.31, and so by setting $G_1 = [G, G]$ we get the largest possible abelian quotient G/G_1 .

Next, we want to choose G_2 so that $G_1/G_2 \leq Z(G/G_2)$, and again we might as well aim for $G_1/G_2 = Z(G/G_2)$. What we want, therefore, is a normal subgroup $G_2 \triangleleft G$ such that the cosets in G_1/G_2 are exactly those that commute with all of the cosets in G/G_2 . We can achieve this by setting $G_2 = [G, G_1] = \{ghg^{-1}h^{-1} : g \in G, h \in G_1\}$. More generally, this process leads to the following descending series for G :

Definition 7.53 For a group G , let $\Gamma_1(G) = G$, and recursively define

$$\Gamma_i(G) = [G, \Gamma_{i-1}(G)] = \{ghg^{-1}h^{-1} : g \in G, h \in \Gamma_{i-1}(G)\}$$

for $i > 1$. The descending series

$$G_n \triangleleft G_{n-1} \cdots \triangleleft G_0 = G,$$

where $G_i = \Gamma_{i+1}(G)$ for $0 \leq i \leq n$, is the **descending central series** or **lower central series** of G .

Let's try this for the dihedral group D_4 again.

Example 7.54 First we set $G_0 = G = D_4$. Next, we set $G_1 = \Gamma_2(G) = [G, G]$, which is $\{e, r^2\}$. The third group $G_2 = \Gamma_3(G) = [G_1, G]$ is the trivial subgroup $\{e\}$, and so the lower central series for D_4 is

$$\{e\} \triangleleft \{e, r^2\} \triangleleft D_4.$$

Earlier we saw that the upper central series for the symmetric group S_3 doesn't reach S_3 , and it turns out that the lower central series doesn't terminate at the trivial subgroup either. To see this, we first set $G_0 = \Gamma_1(S_3) = S_3$. The next group in the series is $G_1 = \Gamma_2(S_3) =$

$[S_3, S_3] = A_3$. And here everything grinds to a halt, because $\Gamma_3(S_3) = [S_3, A_3] = A_3$ as well, so the lower central series for S_3 is therefore

$$A_3 \triangleleft S_3.$$

The upper and lower central series for D_4 happen to be the same, but those for S_3 aren't even the same length.

The case where a group G has a finite-length central series which reaches $\{e\}$ at one end, and G at the other, is particularly interesting:

Definition 7.55 Suppose that a group G has a central series

$$\{e\} = G_0 < G_1 < \cdots < G_n = G$$

connecting G with its trivial subgroup $\{e\}$. Then we say that G is **nilpotent**. The smallest possible length n over all such central series is called the **nilpotency class** of G .

The lower central series is in some sense the optimal descending central series: if it reaches $\{e\}$ then it does so faster than any other descending central series.

Proposition 7.56 Suppose that a group G has a finite descending central series

$$\{e\} = G_{n+1} < G_n < \cdots < G_1 = G.$$

Then $\Gamma_i(G) \leq G_i$ for $0 < i \leq n+1$. Furthermore, if G has nilpotency class c , then $n \geq c+1$.

To prove this, we first need the following lemma.

Lemma 7.57 Let G be a group, and suppose that $K \trianglelefteq G$ and $K \trianglelefteq H \leq G$. Then $H/K \leq Z(G/K)$ if and only if $[G, H] \leq K$.

Proof Suppose that $H/K \in Z(G/K)$. This means that every coset in H/K commutes with every coset in G/K . That is, for every $h \in H$ and $g \in G$ we have $(hK)(gK) = (gK)(hK)$. Moreover, $(hK)(gK) = (hg)K$ and $(gK)(hK) = (gh)K$, so

$$(gh)K = (gK)(hK) = (hK)(gK) = (hg)K$$

and hence by Proposition 2.26 it follows that $(hg)(gh)^{-1} \in K$. But $(hg)(gh)^{-1} = hgh^{-1}g^{-1}$ is the commutator $[h, g] \in [H, G]$. Thus every such commutator lies in K and hence $[H, G] \leq K$.

Each step in this argument is reversible: suppose that $[H, G] \leq K$. Then for any $hgh^{-1}g^{-1} \in [H, G] \leq K$ we have $(hg)K = (gh)K$ and hence $H/K \leq Z(G/K)$. $\square$

Proof of Proposition 7.56 We will prove the inclusion $\Gamma_i(G) \leq G_i$ by induction on i . The inclusion obviously holds for $i = 1$, since $\Gamma_1(G) = G = G_1$.

Now suppose that $\Gamma_i(G) \leq G_i$ for some $i > 1$. Then

$$\Gamma_{i+1}(G) = [G, \Gamma_i(G)] \leq [G, G_i] \leq G_{i+1}.$$

The second of these inclusions follows from Lemma 7.57 by setting $H = G_i$ and $K = G_{i+1}$.

If G has nilpotency class c , then any central series for G must have length at least c , hence $n \geq c+1$. $\square$

What this means is that if G is nilpotent, that is, if there exists a central series of finite length which has G at one end and $\{e\}$ at the other, then the lower central series Γ is finite and terminates at $\{e\}$ as well. (This last statement follows because $\Gamma_{n+1}(G) \leq G_{n+1} = \{e\}$.) Not only that, the lower central series is the shortest descending central series that does this.

An analogous statement holds for the upper central series too:

Proposition 7.58 *Suppose that a group G has a finite ascending central series*

$$\{e\} = G_0 < G_1 < \cdots < G_n = G.$$

Then $G_i \leq Z^i(G)$ for $0 \leq i \leq n$. Furthermore, if G has nilpotency class c , then $n \geq c$.

Proof As before, we will prove the inclusion $G_i \leq Z^i(G)$ by induction on i . The base case is clear, since $G_i = \{e\} = Z^0(G)$.

Now suppose that $G_i \leq Z^i(G)$ for some $i > 0$. For any $g \in G$ and $h \in G_{i+1}$ it follows that $(gG_i)(hG_i) = (hG_i)(gG_i)$ in G/G_i , since $G_{i+1}/G_i \leq Z(G/G_i)$. And since $G_i \leq Z^i(G)$ we also have

$$(gZ^i(G))(hZ^i(G)) = (hZ^i(G))(gZ^i(G)).$$

Since $g \in G$, this means that every coset of the form $hZ^i(G)$ commutes with every coset of the form $gZ^i(G)$, and hence $hZ^i(G) \in Z(G/Z^i(G))$. But $Z(G/Z^i(G)) = Z^{i+1}(G)/Z^i(G)$, and therefore $h \in Z^{i+1}(G)$. We originally chose $h \in G_{i+1}$, hence $G_{i+1} \leq Z^{i+1}(G)$ as claimed.

If G has nilpotency class c , then any central series for G must have length at least c , hence $n \geq c$. $\square$

What all this tells us is that if G is nilpotent, then the upper central series Z is finite and terminates at G . (This is because $G = G_n \leq Z^n(G) \leq G$.) Moreover, the upper central series is the fastest ascending central series that terminates at G .

Corollary 7.59 *The following statements are equivalent:*

- (i) G is nilpotent,
- (ii) $\Gamma_n(G) = \{e\}$ for some $n > 0$, and
- (iii) $Z^n(G) = G$ for some $n \geq 0$.

The groups $\Gamma_i(G)$ in the lower central series are defined recursively in terms of commutator subgroups $\Gamma_{i+1}(G) = [\Gamma_i(G), \Gamma_i(G)]$, and the groups $G^{(i)}$ in the derived series are defined recursively as commutator subgroups $G^{(i+1)}(G) = [G^{(i)}, G^{(i)}]$. The next proposition tells us how these series are related.

Proposition 7.60 *For any group G , $G^{(i)} \leq \Gamma_{i+1}(G)$ for all $i \geq 0$.*

Proof We prove this by induction on i . The base case $i = 0$ holds, since $G^{(0)} = G = \Gamma_1(G)$. Now suppose that $G^{(i)} \leq \Gamma_{i+1}(G)$ for some $i > 0$. Then

$$G^{(i+1)} = [G^{(i)}, G^{(i)}] \leq [G^{(i)}, G] \leq [\Gamma_{i+1}(G), G] = \Gamma_{i+2}(G).$$

Hence $G^{(i)} \leq \Gamma_{i+1}(G)$ for all $i \geq 0$, as claimed. $\square$

Corollary 7.61 *Nilpotent groups are soluble.*

Proof Suppose G is a nilpotent group. Then its lower central series terminates at $\{e\}$ after finitely many steps. Since $\Gamma^{(i)} \leq \Gamma_{i+1}(G)$ for all $i \geq 0$, it follows that the derived series must also reach $\{e\}$ after finitely many steps, and hence by Proposition 7.47 G must be soluble. $\square$

Not all soluble groups are nilpotent, however. For example, S_3 is soluble, because its derived series

$$\{\iota\} \triangleleft A_3 \triangleleft S_3$$

reaches $\{\iota\}$ after finitely many steps. But its lower central series

$$A_3 \triangleleft S_3$$

doesn't reach $\{\iota\}$, so it isn't nilpotent.

We saw earlier that subgroups, quotients, extensions and direct products of soluble groups are soluble. Some analogous results hold for nilpotent groups. Firstly, we show that subgroups of nilpotent groups are nilpotent

Proposition 7.62 *Let G be nilpotent, and suppose that $H \leq G$. Then H is also nilpotent.*

Proof We prove by induction that $\Gamma_i(H) \leq \Gamma_i(G)$. The base case $i = 1$ holds, since $\Gamma_1(H) = H \leq G = \Gamma_1(G)$. Now suppose that $\Gamma_i(H) \leq \Gamma_i(G)$ for some $i > 1$. Then

$$\Gamma_{i+1}(H) = [\Gamma_i(H), H] \leq [\Gamma_i(H), G] \leq [\Gamma_i(G), G] = \Gamma_{i+1}(G).$$

So $\Gamma_i(H) \leq \Gamma_i(G)$ for all $i > 0$. If G is nilpotent then there exists some $n > 0$ such that $\Gamma_n(G) = \{e\}$, and thus $\Gamma_n(H) \leq \Gamma_n(G) = \{e\}$ must also be trivial, so H is nilpotent. $\square$

Next, we show that finite direct products of nilpotent groups are nilpotent.

Proposition 7.63 *If G and H are nilpotent groups, then their direct product $G \times H$ is nilpotent as well.*

Proof We prove this by showing that $\Gamma_i(G \times H) \leq \Gamma_i(G) \times \Gamma_i(H)$ for $i > 0$. Again, we prove this by induction. The base case $i = 1$ holds, since $\Gamma_1(G \times H) = G \times H = \Gamma_1(G) \times \Gamma_1(H)$.

Now suppose that $\Gamma_i(G \times H) \leq \Gamma_i(G) \times \Gamma_i(H)$ for some $i > 1$. Then

$$\begin{aligned} \Gamma_{i+1}(G \times H) &= [\Gamma_i(G \times H), G \times H] \leq [\Gamma_i(G) \times \Gamma_i(H), G \times H] \\ &\leq [\Gamma_i(G), G] \times [\Gamma_i(H), H] = \Gamma_{i+1}(G) \times \Gamma_{i+1}(H), \end{aligned}$$

and so the induction holds for all $i > 1$. Since G and H are nilpotent, there exists $m, n \in \mathbb{N}$ such that $\Gamma_m(G) = \Gamma_n(H) = \{e\}$. Let $k = \max(m, n)$. Then

$$\Gamma_k(G \times H) \leq \Gamma_k(G) \times \Gamma_k(H) = \{e\} \times \{e\} = \{(e, e)\}.$$

Hence the lower central series for $G \times H$ terminates at the trivial subgroup $\{(e, e)\}$ after finitely many steps, so $G \times H$ is nilpotent. $\square$

We can extend this by induction to show that direct products of finitely many nilpotent groups are nilpotent:

Corollary 7.64 *Let $G_1, \dots, G_n$ be nilpotent. Then the direct product $G_1 \times \dots \times G_n$ is nilpotent too.*

Finally, we show that images of nilpotent groups are nilpotent, and consequently so are their quotients.

Proposition 7.65 *Let $f: G \rightarrow H$ be a homomorphism from a nilpotent group G to some group H . Then the image $f(G)$ is nilpotent too.*

Proof Again, we prove this by considering the lower central series, in this case by showing inductively that $\Gamma_i(f(G)) = f(\Gamma_i(G))$ for all $i > 0$.

The base case $i = 1$ holds, since $\Gamma_1(f(G)) = f(G) = f(\Gamma_1(G))$. Now suppose that $f(\Gamma_i(G)) = \Gamma_i(f(G))$ for some $i > 1$, and let $g \in G$ and $h \in \Gamma_i(G)$. Then

$$\begin{aligned} f([g, h]) &= [f(g), f(h)] \in [f(G), f(\Gamma_i(G))] \\ &= [f(G), \Gamma_i(f(G))] = \Gamma_{i+1}(f(G)). \end{aligned}$$

Hence $f(\Gamma_{i+1}(G)) \leq \Gamma_{i+1}(f(G))$.

Now suppose that $x \in f(G)$ and $y \in \Gamma_i(f(G)) = f(\Gamma_i(G))$. Since f maps G surjectively onto $f(G)$, there exists $g \in G$ and $h \in \Gamma_i(G)$ such that $x = f(g)$ and $y = f(h)$. Then

$$[x, y] = [f(g), f(h)] = f([g, h]) \in f([G, \Gamma_i(G)]) = f(\Gamma_{i+1}(G)).$$

Hence $\Gamma_{i+1}(f(G)) \leq f(\Gamma_{i+1}(G))$, so $\Gamma_{i+1}(f(G)) = f(\Gamma_{i+1}(G))$, and inductively $\Gamma_i(f(G)) = f(\Gamma_i(G))$ for all $i > 0$.

Now, since G is nilpotent, there exists some positive integer n such that $\Gamma_n(G) = \{e\}$, and therefore $\Gamma_n(f(G)) = f(\Gamma_n(G)) = f(\{e\}) = \{e\}$, which means that $f(G)$ is nilpotent. $\square$

Corollary 7.66 *If G is a nilpotent group, and $N \trianglelefteq G$, then the quotient G/N is also nilpotent.*

Proof This follows from Proposition 7.65 by considering the canonical surjective homomorphism $q: G \rightarrow G/N$. $\square$

Extensions of nilpotent groups, however, need not be nilpotent. That is, given a group G and normal subgroup $N \trianglelefteq G$ such that N and G/N are nilpotent, it isn't true in general that G must also be nilpotent. (The analogous statement for soluble groups is true, by Proposition 7.40.) The dihedral group D_3 is a simple counterexample: we know D_3 isn't nilpotent, because it's isomorphic to S_3 , which we saw earlier isn't nilpotent. However, it has a normal subgroup $R_3 \cong \mathbb{Z}_3$ which is nilpotent, and the quotient $D_3/R_3 \cong \mathbb{Z}_2$ is also nilpotent.

The nilpotency class (the minimal length of a central series for the group) gives us a way of partially classifying groups. The trivial group $\{e\}$ is the only group with nilpotency class 0. The groups with nilpotency class 1 are exactly the abelian groups: the lower central series for an abelian group A is given by $\Gamma_1(A) = A$ and $\Gamma_2(A) = [A, A] = \{e\}$, so the nilpotency class of A is 1; conversely if A has nilpotency class 1 then its lower central series must terminate with $\Gamma_2(A) = [A, A] = \{e\}$, which only happens if A is abelian.

Proposition 7.43 showed that all finite p -groups are soluble. An analogous result holds with respect to nilpotency:

Proposition 7.67 *Let G be a finite p -group of order $|G| = p^k$ for some prime p and positive integer k . Then G is nilpotent.*

Proof By Proposition 7.19 finite p -groups have nontrivial centres, so $Z^1(G) = Z(G)$ has order a nonzero power of p by Lagrange's Theorem²². Furthermore, $Z^2(G)$ is defined such that $Z^2(G)/Z^1(G) = Z(G/Z^1(G))$. The quotient $G/Z^1(G)$ must be either trivial or a p -group itself. If the former, $Z^1(G) = G$ and hence G is nilpotent by Corollary 7.59; if the former, $Z(G/Z^1(G))$ must be nontrivial by Proposition 7.19, and hence have order a power of p by Lagrange's Theorem. Thus $Z^2(G)$ must be a nontrivial p -group that strictly contains $Z^1(G)$. Proceeding in this way, we find that $Z^i(G) < Z^{i+1}(G)$ for all $i \geq 0$, and since $|G|$ is finite, the upper central series must eventually terminate at G . Hence G is nilpotent by Corollary 7.59. $\square$

Putting this together with Proposition 7.63 and Corollary 7.64, it follows that any direct product of finitely many finite p -groups must

²² Theorem 2.30, page 54.

also be nilpotent. In particular, if a finite group G happens to be a direct product of its Sylow subgroups, then it will therefore be nilpotent. What isn't so obvious is that the converse also holds: all finite nilpotent groups can be expressed as direct products of their Sylow subgroups.

Proposition 7.68 *A finite group G is nilpotent if and only if it is the direct product of its Sylow subgroups.*

To prove this, we first need the following proposition about normalisers in nilpotent groups.

Proposition 7.69 *Let G be a nilpotent group, and $H < G$ a proper subgroup of G . Then $H < N_G(H)$.*

What this says is that every proper subgroup of a nilpotent group is strictly contained in its normaliser.

Proof Since $H < G$, we have $H < \Gamma_1(G) = G$ and $\{e\} = \Gamma_{n+1}(G) \leq H$, where n is the nilpotency class of G . Now choose $i > 0$ such that $\Gamma_i(G) \leq H < \Gamma_{i-1}(G)$. Then

$$[\Gamma_{i-1}(G), H] \leq [\Gamma_{i-1}(G), G] = \Gamma_i(G) \leq H,$$

which means for any $g \in \Gamma_{i-1}(G)$ and $h \in H$, we have $ghg^{-1}h^{-1} = [g, h] \in H$, and thus $ghg^{-1} \in H$. Hence $gHg^{-1} = H$ for all $g \in \Gamma_{i-1}(G)$, which means that $\Gamma_{i-1}(G) \leq N_G(H)$. And since $H < \Gamma_{i-1}(G)$, it follows that $H < N_G(H)$. $\square$

In the proof of Proposition 7.68, we will use the contrapositive form of this result:

Corollary 7.70 *Let G be a nilpotent group, and suppose that $H \leq G$ such that $H = N_G(H)$. Then $H = G$.*

One half of the proof of Proposition 7.68 is now straightforward, but the other half requires a short lemma about normalisers of Sylow subgroups.

Lemma 7.71 *Let H be a Sylow p -subgroup of a finite group G , for some prime p . Then $N_G(N_G(H)) = N_G(H)$.*

Proof By the definition of the normaliser²³, $H \trianglelefteq N_G(H)$, and by Corollary 7.12, H must be the unique Sylow p -subgroup of $N = N_G(H)$. For any $g \in N_G(N)$ we have $gHg^{-1} \leq gNg^{-1} = N$, so gHg^{-1} is also a Sylow p -subgroup of N and therefore $gHg^{-1} = H$. This means that g normalises H in G , so $g \in N_G(H) = N$. Hence $N_G(N) \triangleleft N$, and thus $N_G(N) = N_G(N_G(H)) = N_G(H)$ as claimed. $\square$

²³ Definition 4.46, page 120.

We now have the necessary ingredients to prove our result about finite nilpotent groups.

Proof of Proposition 7.68 Suppose that G is the direct product of its

Sylow subgroups. Each Sylow subgroup is a p -group for some prime p , and hence nilpotent by Proposition 7.67. The direct product of finitely many nilpotent groups is nilpotent by Proposition 7.63 and Corollary 7.63, so G is nilpotent.

To prove the converse, suppose that G is finite and nilpotent, and that H is a Sylow p -subgroup of G for some prime p . Let $N = N_G(H)$ be the normaliser of H in G . By Lemma 7.71 we know that $N_G(N) = N$; that is, N is self-normalising in G . By Corollary 7.70 $N = N_G(H) = G$, and therefore $H \trianglelefteq G$.

Now suppose that $|G| = p_1^{n_1} \cdots p_k^{n_k}$ for some positive integer k , where $p_1, \dots, p_k$ are distinct primes, and $n_1, \dots, n_k \in \mathbb{N}$. By Sylow's First Theorem²⁴ we know that G has a Sylow p_i -subgroup H_i of order $p_i^{n_i}$ for $1 \leq i \leq k$, and we have just seen that all of these are normal in G . We want to show that $H_1 \cdots H_k \cong H_1 \times \cdots \times H_k$, and we prove this by induction, showing that $H_1 \cdots H_i \cong H_1 \times \cdots \times H_i$ for $1 \leq i \leq k$. The base case $i = 1$ is clearly true; suppose that the proposition is true for some $i > 1$.

We now consider $(H_1 \times \cdots \times H_i)H_{i+1}$. The first part of this has order $|H_1 \times \cdots \times H_i| = p_1^{n_1} \cdots p_i^{n_i}$, while the second has order $|H_{i+1}| = p_{i+1}^{n_{i+1}}$. By Proposition 2.34, the order of any element of $H_1 \times \cdots \times H_i$ must divide $p_1^{n_1} \cdots p_i^{n_i}$, and any element of H_{i+1} must divide $p_{i+1}^{n_{i+1}}$. These are coprime, so any element in $(H_1 \times \cdots \times H_i) \cap H_{i+1}$ must have order 1, and therefore by Proposition 1.19 $(H_1 \times \cdots \times H_i) \cap H_{i+1} = \{e\}$. By Corollary 3.23, $(H_1 \times \cdots \times H_i)H_{i+1}$ must be the internal direct product of $H_1, \dots, H_{i+1}$ and hence the inductive step holds.

In particular, setting $i = k$ we see that $G = H_1 \cdots H_k \cong H_1 \times \cdots \times H_k$ as claimed. $\square$

In the middle of this proof we also derived the following useful fact:

Corollary 7.72 *A finite group G is nilpotent if and only if all of its Sylow subgroups are normal.*

We will now briefly look at a third class of groups defined in terms of subgroup series. By Definition 7.49, a group G is nilpotent if there exists a finite-length central series stretching between $\{e\}$ and G . Recall that for a series to be central, we require $G_{i+1}/G_i \leq Z(G/G_i)$, and also that the series is normal; that is, $G_i \trianglelefteq G$ for all i .

More generally, a subnormal series is abelian if each quotient G_{i+1}/G_i is abelian (but not necessarily contained in the centre of G/G_i); a group has a finite-length abelian series if and only if it is soluble.

Between these two, we can consider the case where G has a finite-length normal series where the quotients G_{i+1}/G_i are cyclic.

²⁴ Theorem 7.4, page 236.

Definition 7.73 A group G is **supersoluble** or **supersolvable** if there exists a finite-length normal series

$$\{e\} = G_0 < G_1 < \cdots < G_n = G$$

such that each quotient G_{i+1}/G_i is cyclic.

Since every normal series is a subnormal series, a supersoluble group has a finite-length subnormal series with cyclic factors. This can be refined to a finite-length subnormal series whose factors are cyclic groups of prime order, and hence every supersoluble group is soluble. There exist soluble groups that aren't supersoluble, however:

Example 7.74 The symmetric group S_4 is soluble by the existence of the subnormal series

$$\{1\} \triangleleft \langle (12)(34) \rangle \triangleleft \langle (12)(34), (13)(24) \rangle \triangleleft A_4 \triangleleft S_4,$$

the factors of which are isomorphic to $\mathbb{Z}_2$, $\mathbb{Z}_2$, $\mathbb{Z}_3$ and $\mathbb{Z}_2$. However, this series is not normal since $\langle (12)(34), (13)(24) \rangle \not\triangleleft S_4$. Moreover, there is no normal series for S_4 with cyclic factors, since the only normal subgroups of S_4 are $\{1\}$, $\langle (12)(34), (13)(24) \rangle \cong V_4$ and A_4 , from which we can't construct a normal series with only cyclic factors. The alternating group A_4 is also soluble but not supersoluble.

Results analogous to Propositions 7.38, 7.39, 7.40 and Corollary 7.41 hold for supersoluble groups:

Proposition 7.75 *If G is a supersoluble group, and $H < G$ is a subgroup of G , then H is supersoluble.*

Proof The proof is very similar to that of Proposition 7.38, together with the observation that if $H_{i+1}/H_i < G_{i+1}/G_i$ and G_{i+1}/G_i is cyclic then so is H_{i+1}/H_i by Proposition 2.8. $\square$

Proposition 7.76 *If G is a supersoluble group and $N \trianglelefteq G$ is normal in G , then the quotient G/N is also supersoluble.*

Proof Again, the result follows by an almost identical argument to that used in the proof of Proposition 7.39. $\square$

Proposition 7.77 *If G and H are supersoluble groups, then so is their direct product $G \times H$.*

Proof Suppose that G has a cyclic normal series

$$\{e\} = G_0 < G_1 < \cdots < G_m = G$$

and that H has a cyclic normal series

$$\{e\} = H_0 < H_1 < \cdots < H_n = H.$$

Then for $0 \leq i < m$ we have $G_{i+1} \times H_0 / G_i \times H_0 \cong G_{i+1}/G_i$, which is cyclic since G is supersoluble. Similarly, for $0 \leq i < n$, it follows that

$G_m \times H_{i+1} / G_m \times H_i \cong H_{i+1} / H_i$ is also cyclic since H is supersoluble. Therefore the series

$$\begin{aligned} \{e\} = G_0 \times H_0 &< G_1 \times H_0 < \cdots < G_m \times H_0 \\ &< G_m \times H_1 < \cdots < G_m \times H_n = G \times H \end{aligned}$$

is a cyclic normal series for $G \times H$, and hence the direct product $G \times H$ is supersoluble. $\square$

By induction it follows that a direct product of finitely many supersoluble groups is also supersoluble.

The analogue of Proposition 7.40 doesn't hold in general: the Klein group $V_4 \cong \langle (12)(34), (13)(24) \rangle = N$ is supersoluble and the quotient $A_4/N \cong \mathbb{Z}_3$ is supersoluble as well, but A_4 isn't. However, a slightly weaker result does hold:

Proposition 7.78 *Let $N \triangleleft G$ be a normal subgroup of some group G . If N is cyclic and G/N is supersoluble, then G is supersoluble.*

Proof If G/N is supersoluble, then there exists a cyclic normal series

$$\{e\} = H_1 < H_1 < \cdots < H_n = G/N.$$

Each group H_i , for $1 \leq i \leq n$, can be expressed as a quotient G_i/N for some other group $G_i \triangleleft G$, so our cyclic normal series can be rewritten

$$N/N = G_1/N < G_1/N < \cdots < G_n/N = G/N,$$

which we can then use to form a new series

$$\{e\} = G_0 < G_1 < \cdots < G_n = G,$$

by inserting the trivial subgroup $G_0 = \{e\}$ at the beginning. Each factor G_{i+1}/G_i of this new series is cyclic, since each factor H_{i+1}/H_i is cyclic, and $H_{i+1}/H_i = (G_{i+1}/N)/(G_i/N) \cong G_{i+1}/G_i$ by the Third Isomorphism Theorem.²⁵ Therefore G also has a cyclic normal series and is thus supersoluble. $\square$

We now have a hierarchy of classes of finite groups:

$$\text{trivial} \subset \text{cyclic} \subset \text{abelian} \subset \text{nilpotent} \subset \text{supersoluble} \subset \text{soluble}$$

²⁵ Theorem 4.65, page 128.

If you have built castles in the air, your work need not be lost; that is where they should be. Now put the foundations under them.

— Henry Thoreau (1817–1862),
Walden (1854) 346

7.4 Semidirect products

THE INTERNAL DIRECT PRODUCT enables us to decompose a group as the direct product of two normal subgroups. More precisely, if H and K are both normal subgroups of a group G with trivial intersection $H \cap K = \{e\}$, and if $HK = G$, then G is isomorphic to the direct product

$H \times K$. Also, we have $G/H \cong K$ and $G/K \cong H$. But what happens if we only require one of H and K to be normal?

Suppose that $H \trianglelefteq G$ and $K \leq G$ with $H \cap K = \{e\}$ and $HK = G$. What groups G satisfy this set of conditions for two given subgroups H and K ? Equivalently, given two groups H and K , what groups G exist such that $H \trianglelefteq G$ and $K \leq G$ with $H \cap K = \{e\}$ and $G = HK$?

For example, suppose $H \cong \mathbb{Z}_3$ and $K \cong \mathbb{Z}_2$. Up to isomorphism there are two ways of combining these to form a new group satisfying the above properties. The obvious one is the direct sum $\mathbb{Z}_3 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_6$. Here $H = \{0, 2, 4\}$ and $K = \{0, 3\}$. Then $H \cap K = \{0\}$ and $H + K = \mathbb{Z}_6$. Furthermore, $H \trianglelefteq \mathbb{Z}_6$ and $K \leq \mathbb{Z}_6$ (actually $K \trianglelefteq \mathbb{Z}_6$ as well, but we don't mind), and $\mathbb{Z}_6/H \cong \mathbb{Z}_2 \cong K$.

The other, slightly less obvious way yields $D_3 \cong S_3$. Set $H = R_3 = \{e, r, r^2\}$ and $K = \{e, m_1\}$. Then $H \trianglelefteq D_3$ and $K \leq D_3$ with $H \cap K = \{e\}$ and $HK = D_3$. Furthermore, $D_3/H \cong \mathbb{Z}_2 \cong K$.

To study this idea in more detail, we introduce the following definition:

Definition 7.79 Let G be a group, and suppose that $H \trianglelefteq G$ is normal in G and $K \leq G$ is a subgroup of G . If $H \cap K = \{e\}$ and $G = HK$ then we say that G is an **(internal) semidirect product** of H by K .

We'll meet the external semidirect product shortly.

For a direct product $G = H \times K$, the case where both H and K are normal in G , we know by Proposition 3.22 that the elements of H commute with those of K ; that is, for any $h \in H$ and $k \in K$, we have $hk = kh$. Or, equivalently, that $khk^{-1} = h$. What this means is that the conjugation action²⁶ of K on H is trivial.

²⁶ Definition 6.11, page 211.

Applying this insight to the more general case of a semidirect product $G = HK$, we find that K acts by conjugation on H . Since H is a normal subgroup, it is closed under conjugation by any element of G , and therefore also by any element of $K \leq G$. The next proposition confirms that this is a well-defined group action of K on H .

Proposition 7.80 Suppose that a group G can be decomposed as a semidirect product $G = HK$, with $H \trianglelefteq G$. Then K acts on H by conjugation.

Proof For any $h \in H$ and $k \in K$ we define $k \cdot h = khk^{-1}$. This lies in H because H is normal and hence closed under conjugation by any element of G , and in particular by any element of $K \leq G$.

Clearly $e \cdot h = ehe^{-1} = h$, so the identity element $e \in K$ acts trivially on H . Also, for any $k_1, k_2 \in K$ we have

$$k_1 \cdot (k_2 \cdot h) = k_1 \cdot (k_2 h k_2^{-1}) = k_1 k_2 h k_2^{-1} k_1^{-1} = (k_1 k_2) h (k_1 k_2)^{-1} = (k_1 k_2) \cdot h,$$

and hence this is a well-defined action of K on H . □

	e	r	r^2	m_1	m_2	m_3
e	e	e	e	e	e	e
r	r	r	r	r^2	r^2	r^2
r^2	r^2	r^2	r^2	r	r	r
m_1	m_1	m_2	m_3	m_1	m_3	m_2
m_2	m_2	m_3	m_1	m_3	m_2	m_1
m_3	m_3	m_1	m_2	m_2	m_1	m_3

Table 7.2: Conjugation in D_3

Applying this to the case of D_3 , with $H = R_3 = \{e, r, r^2\}$ and $K = \{e, m_1\}$, and considering the conjugation action of K on H (see Table 7.2) we find that the conjugation action of K on H maps $h \mapsto h^{-1}$ for any $h \in H$. That is, $k \cdot h = khk^{-1} = h^{-1}$ for all $k \in K$ and $h \in H$. We can use this to construct a presentation for HK :

$$\langle r, m_1 : r^3 = e, m_1^2 = e, m_1 r m_1^{-1} = r^{-1} \rangle$$

This presentation is the $n=3$ case of the presentation for D_n obtained in Proposition 5.28.

There are only two ways that $\mathbb{Z}_2$ can act on $\mathbb{Z}_3$: this one, and the trivial action, which corresponds to the direct sum $\mathbb{Z}_3 \oplus \mathbb{Z}_2$.

So, if a group G decomposes as an internal semidirect product $G = HK$, then as part of this we automatically get an action of K on H . Conversely, two groups H and K , together with a specified action of K on H , can be combined to form a unique group $G = HK$:

Proposition 7.81 *Let H and K be groups, with K acting on H via an action $\phi: K \rightarrow \text{Aut}(H)$. The multiplication operation*

$$(h_1, k_1) * (h_2, k_2) = (h_1 \phi_{k_1}(h_2), k_1 k_2) = (h_1(k_1 \cdot h_2), k_1 k_2)$$

defines a group structure on the Cartesian product $H \times K$. The resulting group $G = H \rtimes K$ decomposes as an internal semidirect product of $H \times \{e_K\}$ by $\{e_H\} \times K$. The $(\{e_H\} \times K)$ -action on $H \times \{e_K\}$ is determined by ϕ .

Proof To see that the multiplication operation is associative, suppose $h_1, h_2, h_3 \in H$ and $k_1, k_2, k_3 \in K$. Then

$$\begin{aligned} ((h_1, k_1) * (h_2, k_2)) * (h_3, k_3) &= (h_1 \phi_{k_1}(h_2), k_1 k_2) * (h_3, k_3) = \\ &= (h_1 \phi_{k_1}(h_2) \phi_{k_1 k_2}(h_3), k_1 k_2 k_3) = (h_1 \phi_{k_1}(h_2 \phi_{k_2}(h_3)), k_1 k_2 k_3) = \\ &= (h_1, k_1) * (h_2 \phi_{k_2}(h_3), k_2 k_3) = (h_1, k_1) * ((h_2, k_2) * (h_3, k_3)). \end{aligned}$$

The identity element is (e_H, e_K) since, for any $h \in H$ and $k \in K$,

$$(e_H, e_K) * (h, k) = (e_H \phi_{e_K}(h), e_K k) = (h, k)$$

$$\text{and } (h, k) * (e_H, e_K) = (h \phi_k(e_H), k e_K) = (h, k).$$

The inverse of (h, k) is $(\phi_{k^{-1}}(h^{-1}), k^{-1})$, since

$$\begin{aligned} (h, k) * (\phi_{k^{-1}}(h^{-1}), k^{-1}) &= (h \phi_k(\phi_{k^{-1}}(h^{-1})), k k^{-1}) = \\ &= (h \phi_{k k^{-1}}(h^{-1}), e_K) = (h \phi_{e_K}(h^{-1}), e_K) = (h h^{-1}, e_K) = (e_H, e_K) \end{aligned}$$

and

$$\begin{aligned} (\phi_{k^{-1}}(h^{-1}), k^{-1}) * (h, k) &= (\phi_{k^{-1}}(h^{-1}) \phi_{k^{-1}k}(h), k^{-1}k) = \\ &= (\phi_{k^{-1}}(h^{-1}h), e_K) = (\phi_{k^{-1}}(e_H), e_K) = (e_H, e_K). \end{aligned}$$

Therefore the set $H \times K$ forms a group G with this multiplication operation.

There are injective maps $f_H: H \hookrightarrow G$ and $f_K: K \hookrightarrow G$ defined by $f_H(h) = (h, e_K)$ and $f_K(k) = (e_H, k)$. These are homomorphisms, since

$$f_H(h_1 h_2) = (h_1 h_2, e_K) = (h_1, e_K) * (h_2, e_K) = f_H(h_1) * f_H(h_2)$$

$$\text{and } f_K(k_1 k_2) = (e_H, k_1 k_2) = (e_H, k_1) * (e_H, k_2) = f_K(k_1) * f_K(k_2),$$

so $\text{im}(f_H) = H \times \{e_K\}$ and $\text{im}(f_K) = \{e_H\} \times K$ are subgroups.

Furthermore, $\text{im}(f_H)$ is normal in G , since for any $h_1, h_2 \in H$ and $k \in K$ we have

$$\begin{aligned} (h_2, k) * (h_1, e_K) * (h_2, k)^{-1} &= (h_2, k) * (h_1, e_K) * (\phi_{k^{-1}}(h_2^{-1}), k^{-1}) \\ &= (h_2 \phi_k(h_1), k) * (\phi_{k^{-1}}(h_2^{-1}), k^{-1}) = (h_2 \phi_k(h_1) \phi_k(\phi_{k^{-1}}(h_2^{-1})), k k^{-1}) \\ &= (h_2 \phi_k(h_1) h_2^{-1}, e_K) \in H \times \{e_K\}, \end{aligned}$$

and so $\text{im}(f_H)$ is closed under conjugation by any element of G . Clearly $(H \times \{e_K\}) \cap (\{e_H\} \times K) = \{e_H, e_K\}$ and $(H \times \{e_K\})(\{e_H\} \times K) = G$, so G decomposes as an internal semidirect product, as claimed.

Finally, given $h \in H$ and $k \in K$,

$$\begin{aligned} (e_H, k) * (h, e_K) * (e_H, k)^{-1} &= (e_H, k) * (h, e_K) * (e_H, k^{-1}) \\ &= (\phi_k(h), k k^{-1}) = (\phi_k(h), e_K), \end{aligned}$$

and hence the conjugation action of $\{e_H\} \times K$ on $H \times \{e_K\}$ is determined completely by ϕ , the K -action on H . $\square$

Definition 7.82 Let H and K be groups, and $\phi: K \rightarrow \text{Aut}(H)$ be an action of K on H . Then the group G from Proposition 7.81 is called the **(external) semidirect product** of H by K , and is denoted $H \rtimes K$,²⁷ or sometimes $H \rtimes_{\phi} K$. We say that $H \rtimes K$ **realises** the action ϕ .

Example 7.83 The dihedral group D_n can be constructed as a semidirect product $\mathbb{Z}_n \rtimes \mathbb{Z}_2$. There are two $\mathbb{Z}_2$ -actions we can impose on $\mathbb{Z}_n$: the trivial action, and the action $\phi: \mathbb{Z}_2 \rightarrow \text{Aut}(\mathbb{Z}_n)$ defined by $\phi_0(i) = i$ and $\phi_1(i) = n - i$ for $0 \leq i < n$. The trivial action gives

$$\langle x, y : x^n = 1, y^n = 1, yxy^{-1} = x \rangle$$

which is isomorphic to $\mathbb{Z}_n \oplus \mathbb{Z}_2$ by Proposition 5.22. The nontrivial action ϕ yields the presentation

$$\langle x, y : x^n = 1, y^n = 1, yxy^{-1} = x^{-1} \rangle$$

which is isomorphic to D_n by Proposition 5.28.

Example 7.84 The symmetric group S_n can be constructed as a semidirect product $A_n \rtimes \mathbb{Z}_2$. Recall that $A_n \triangleleft S_n$ is the normal subgroup of even permutations on n objects. Represent $\mathbb{Z}_2$ as the order-2 subgroup $K = \langle (12) \rangle = \{1, (12)\}$ in S_n . Clearly $A_n K = S_n$, since composing all the even permutations in A_n with (12) yields all the

²⁷ Think of the symbol $\rtimes$ as a hybrid of $\times$ and $\triangleleft$: the notation $G = H \rtimes K$ tells us that H is the normal subgroup. Similarly, an expression of the form $G = H \rtimes K$ denotes that K is normal.

odd permutations in $S_n \setminus A_n$. Furthermore, $A_n \cap K = \{\iota\}$, so S_n is the internal semidirect product of A_n with K .

We can form S_n as the external semidirect product $A_n \rtimes \mathbb{Z}_2$ via the action $\phi: \mathbb{Z}_2 \rightarrow \text{Aut}(A_n)$ given by $\phi_0(\sigma) = \sigma$ and $\phi_1(\sigma) = (12)\sigma(12)$ for any $\sigma \in A_n$.

Example 7.85 Let $\mathbb{R}^n$ denote the additive group of $(\mathbb{R}^n, +)$; that is, the set of ordered n -tuples of real numbers, or n -component real vectors. The group $GL_n(\mathbb{R})$ acts on $\mathbb{R}^n$ by matrix multiplication, and hence we can form the semidirect product $\mathbb{R}^n \rtimes GL_n(\mathbb{R})$.

The underlying set of this group is $\mathbb{R}^n \times GL_n(\mathbb{R})$: the set of ordered pairs $(\mathbf{u}, A)$, with $\mathbf{u}$ a vector in $\mathbb{R}^n$ and A a nonsingular $n \times n$ matrix in $GL_n(\mathbb{R})$. The product of two elements $(\mathbf{u}, A)$ and $(\mathbf{v}, B)$ is

$$(\mathbf{u}, A) * (\mathbf{v}, B) = (\mathbf{u} + A\mathbf{v}, AB).$$

This is isomorphic to the **affine general linear group** $AGL_n(\mathbb{R})$ of transformations $\mathbf{w} \mapsto A\mathbf{w} + \mathbf{u}$ for $A \in GL_n(\mathbb{R})$ and $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$.

For example, suppose that $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and $A, B \in GL_n(\mathbb{R})$. Then $(\mathbf{v}, B)$ acts on $\mathbf{w}$ by $\mathbf{w} \mapsto B\mathbf{w} + \mathbf{v}$ and $(\mathbf{u}, A)$ acts on $\mathbf{w}$ by $\mathbf{w} \mapsto A\mathbf{w} + \mathbf{u}$. By composition, we find that

$$((\mathbf{u}, A)(\mathbf{v}, B))\mathbf{w} = (\mathbf{u}, A)(B\mathbf{w} + \mathbf{v}) = AB\mathbf{w} + A\mathbf{v} + \mathbf{u} = (A\mathbf{v} + \mathbf{u}, AB)\mathbf{w}$$

and hence in $AGL_n(\mathbb{R})$ the multiplication operation is given by

$$(\mathbf{u}, A)(\mathbf{v}, B) = (\mathbf{u} + A\mathbf{v}, AB)$$

which is exactly that given by the semidirect product construction.

An interesting case arises when we take the semidirect product of a group with its full automorphism group.

Example 7.86 Let G be a group, and let its automorphism group $\text{Aut}(G)$ act on it in the obvious way. That is, given an automorphism $\phi \in \text{Aut}(G)$, we define $\phi \cdot g = \phi(g)$. The semidirect product $\text{Hol}(G) = G \rtimes \text{Aut}(G)$ is called the **holomorph** of G .

For example, let $G = \mathbb{Z}_3$. The automorphism group $\text{Aut}(\mathbb{Z}_3) \cong \mathbb{Z}_2$, comprising the identity automorphism and the inverse map. The holomorph $\text{Hol}(\mathbb{Z}_3) \cong \mathbb{Z}_3 \rtimes \mathbb{Z}_2$ is isomorphic to the dihedral group D_3 , or the symmetric group S_3 . (This is the nontrivial case of Example 7.83 where $n = 3$.)

Another important class of semidirect products can be constructed by applying a permutation action to a direct product of copies of a chosen group. Let G be a group, and choose $n \in \mathbb{N}$. Then we can define an action of the symmetric group S_n on the n -fold direct product $G^n = G \times \cdots \times G$, by permuting the coordinates. That is,

given a permutation $\sigma \in S_n$ and an element $(g_1, \dots, g_n) \in G^n$, define $\sigma \cdot (g_1, \dots, g_n)$ to be $(g_{\sigma(1)}, \dots, g_{\sigma(n)})$.

The semidirect product $G^n \rtimes S_n$ obtained via this action is called the **permutation wreath product** of G with S_n , and denoted $G \wr S_n$, or sometimes $G \text{ pwr } S_n$.

More generally, by Cayley's Theorem²⁸ we know that any group H can be regarded as a group of permutations of its underlying set. We can use this to define a wreath product $G \wr H$. Suppose that $|H| = n$ and form the n -fold direct product $G^n = G \times \dots \times G$. Now define the H -action on G^n by using the permutations obtained from Cayley's Theorem: an element of G^n is an n -tuple of elements of G , indexed by elements of $H = \{h_1, \dots, h_n\}$. For any element $h \in H$ we define

²⁸ Theorem 2.13, page 47.

$$h \cdot (g_{h_1}, \dots, g_{h_n}) = (g_{hh_1}, \dots, g_{hh_n}).$$

The corresponding semidirect product $G^n \rtimes H$ is the **regular wreath product** of G by H , and denoted $G \text{ rwr } H$, or $G \wr H$ where the meaning is clear from context.

Example 7.87 The permutation wreath product $\mathbb{Z}_m \wr S_n$ yields the **generalised symmetric group** $S_{m,n}$. If $m = 1$ then this is just the ordinary symmetric group $S_{1,n} = S_n$, while if $m = 2$ then we obtain the **signed symmetric group** $S_{2,n}$ or **hyperoctahedral group** B_n , and if $m = n = 2$ then we get the wreath product $\mathbb{Z}_2 \wr \mathbb{Z}_2$, which is isomorphic to the dihedral group D_4 .

Wreath products have a number of applications in the classification of permutation groups, and in the study of semigroups and automata; however these applications would be too much of a digression here.

7.5 Extensions

WE CAN USE THE SEMIDIRECT PRODUCT to build a wide range of larger groups from smaller ones (or to describe many larger groups as internal semidirect products of their subgroups). Is this all we need? Can we construct or decompose any group in this way?

Unfortunately we can't. For example, $\mathbb{Z}_4$ can't be split as a semidirect product of two smaller subgroups. We can certainly find a normal subgroup of order 2, namely $\{0, 2\} \cong \mathbb{Z}_2$, but there isn't another one. So $\mathbb{Z}_4$ can't be expressed as an internal semidirect product of two subgroups isomorphic to $\mathbb{Z}_2$. Looking at this as an external semidirect product $\mathbb{Z}_2 \rtimes \mathbb{Z}_2$, we discover almost immediately that there is only one $\mathbb{Z}_2$ -action on $\mathbb{Z}_2$: the trivial one. Hence $\mathbb{Z}_2 \rtimes \mathbb{Z}_2 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

No power and no treasure can outweigh the extension of our knowledge.

— Democritus (c.460–c.370BC),
in: John Owen (1833–1896),
Evenings with the Skeptics (1881) I:149

We therefore need to find some way of further generalising the semidirect product. The key is to relax the requirement on the second subgroup even more. With the direct product $G = H \times K$, both H and K embed as normal subgroups of G ; with a semidirect product $G = H \rtimes K$, we only require H to embed as a normal subgroup, while K can be any subgroup satisfying $HK = G$ and $H \cap K = \{e\}$. One option would be to allow H not to be normal in G , alternatively we could weaken the requirement that K embed as a subgroup, and only ask that it form a subset. It is this second approach that we'll adopt. What we're actually trying to do here is understand **group extensions**:

Definition 7.88 Suppose that G , H and K are groups such that there exists an inclusion homomorphism $i: H \hookrightarrow G$ with $i(H) \trianglelefteq G$, and $G/i(H) \cong K$. Then G is an **extension** of K by H .²⁹

²⁹ In some books this is called an **extension** of H by K .

Clearly the direct product $G = H \times K$ satisfies this definition, and the required inclusion is $i_1: H \hookrightarrow H \times K$, mapping H to the first coordinate. The (external) semidirect product $H \rtimes K$ also works; again H is included via $i_1: H \hookrightarrow H \rtimes K$.

What we need to do is look at how the other group K is included. With the direct product $H \times K$ and the semidirect product $H \rtimes K$, we use the inclusion homomorphisms $i_2: K \hookrightarrow H \times K$ or $i_2: K \hookrightarrow H \rtimes K$ mapping K to the second coordinate. If we don't require the image of K to be a subgroup, then we just need a function $s: K \hookrightarrow G$ such that $i(H) \cap s(K) = \{e\}$ and $i(H)s(K) = G$.

Some books use the term **section** to refer only to the case where s is a homomorphism rather than just a function defined on the underlying set.

Definition 7.89 Suppose that G is an extension of a group K by a group H , with inclusion homomorphism $i: H \hookrightarrow G$, and where $i(H) \trianglelefteq G$ and $G/i(H) \cong K$, with quotient homomorphism $q: G \rightarrow K$. Then an inclusion function $s: K \hookrightarrow G$ is called a **lifting** or a **section** if $s(e_K) = e_G$, and if $q \circ s = \text{id}_K$; that is, if $q(s(k)) = k$ for all $k \in K$.

A section $s: K \hookrightarrow G$ determines an element $s(k)$ of the preimage $q^{-1}(k)$ for every element $k \in K$. By Proposition 4.38, these preimages are the cosets of $H = \ker(q)$ in G , and so a section s is the same as a collection of elements of G , one from each coset of H . We actually met something just like this earlier: a section is a special sort of **transversal**.³⁰ We can represent this scenario with the following diagram:

$$H \xhookrightarrow{i} G \begin{array}{c} \xrightarrow{q} \\ \xleftarrow{s} \end{array} K$$

Also, a section $s: K \rightarrow G$ is injective, and so the image $s(K)$ is a subset of G whose elements are in bijective correspondence with the elements of K . If s is a homomorphism then $s(K)$ will be a subgroup of G isomorphic to K , but if not it will just be some distorted copy of K that isn't closed under multiplication. Nevertheless, it will still be the case

³⁰ Definition 5.71, page 185.

that $i(H)s(K) = G$ and $i(H) \cap s(K) = \{e_G\}$ as required.

The section s defines a K -action on H : the images of the elements of K act by conjugation in G on the images of the elements of H . That is,

$$i(k \cdot h) = s(k)i(h)s(k)^{-1} \quad (7.20)$$

for any $h \in H$ and $k \in K$. We can view this as a commutation rule:

$$s(k)i(h) = i(k \cdot h)s(k) \quad (7.21)$$

The semidirect product $G = H \rtimes K$ is the special case where s is a homomorphism; such extensions are said to be **split**, and the section s is sometimes called a **splitting** or **splitting homomorphism**. Our aim now is to develop a classification theory for extensions of some group K by some group H . In general, this is quite complicated, so we will focus on **abelian extensions**: the case where the kernel H is abelian. We also want to define a suitable equivalence relation on extensions, to enable us to classify those extensions that are essentially different.

From now on, suppose G is a group, and A is an abelian group with a predetermined (and possibly trivial) G -action defined on it. Before we go any further, we introduce the following terminology:

Definition 7.90 Let G be a group. A G -**module** is an abelian group A equipped with a G -action $\phi: G \rightarrow \text{Aut}(A)$.

We want to list, up to isomorphism, all groups E such that $E/i(A) \cong G$ for some inclusion homomorphism $i: A \hookrightarrow E$, and where the conjugation $s(G)$ -action on $i(A)$ is determined by the specified G -action on A . That is, we want to classify all extensions of the form

$$A \xhookrightarrow{i} E \twoheadrightarrow^q G \quad (7.22)$$

up to some notion of equivalence.

The definition of equivalence we need must incorporate isomorphism of the middle group E in the diagram (7.22), but in such a way that the inclusion and quotient isomorphisms line up properly:

Definition 7.91 Let G be a group, and let A be a G -module. Two extensions

$$A \xhookrightarrow{i_1} E_1 \twoheadrightarrow^{q_1} G \quad \text{and} \quad A \xhookrightarrow{i_2} E_2 \twoheadrightarrow^{q_2} G$$

are **equivalent** if there exists an isomorphism $f: E_1 \rightarrow E_2$ such that $i_2 = f \circ i_1$ and $q_1 = q_2 \circ f$. That is, if the following diagram commutes:

$$\begin{array}{ccccc} & & E_1 & & \\ & \nearrow i_1 & \downarrow f & \searrow q_1 & \\ A & & & & G \\ & \searrow i_2 & \downarrow q_2 & \nearrow & \\ & & E_1 & & \end{array}$$

Suppose, then, that we have an extension

$$A \xhookrightarrow{i} E \xrightarrow{q} G$$

with a section $s: G \rightarrow E$, and where A has a G -action $\phi: G \rightarrow \text{Aut}(A)$. Choose some element $g \in G$. Then any element $x \in E$ can be represented as a product $x = i(a)s(g)$ for some unique $a \in A$. Furthermore, $s(g)$ acts by conjugation on $i(A)$; from now on we will assume that this $s(G)$ -action on $i(A)$ agrees with the predetermined G -action ϕ . We say that s **realises** the G -action ϕ .

Now suppose that $g, h \in G$ and consider two arbitrary elements $x = i(a)s(g)$ and $y = i(b)s(h)$ in E . Their product

$$\begin{aligned} xy &= i(a)s(g)i(b)s(h) = i(a)s(g)i(b)s(g)^{-1}s(g)s(h) \\ &= i(a)i(\phi_g(b))s(g)s(h) = i(a+\phi_g(b))s(g)s(h). \end{aligned}$$

If s is a homomorphism then $s(g)s(h) = s(gh)$ and this reduces to the semidirect product $A \rtimes G$; but if not, then it doesn't. We can, however, find an element $\sigma(g, h) \in A$ such that

$$s(g)s(h) = s(gh)i(\sigma(g, h)).$$

Doing this for all $g, h \in G$ we get a function $\sigma: G \times G \rightarrow A$. This function, called a **factor set** or **cocycle**, measures how badly s fails to be a homomorphism. The individual elements $\sigma(g, h)$ can be thought of as correction factors or error terms, and if s is a homomorphism, then $\sigma(g, h) = 0$ for all $g, h \in G$. This enables us to define a group structure on the set $A \times G$:

$$(a, g) * (b, h) = (a + \phi_g(b) + \sigma(g, h), gh). \quad (7.23)$$

Not just any function $\sigma: G \times G \rightarrow A$ will do, however: we need (7.23) to determine a valid group structure on the set $A \times G$.

Proposition 7.92 *Let G be a group and A be a G -module. Then a function $\sigma: G \times G \rightarrow A$ is a **factor set** for an extension*

$$A \xhookrightarrow{i} E \xrightleftharpoons[s]{q} G$$

if and only if

$$\sigma(e_G, g) = \sigma(g, e_G) = 0$$

for any $g \in G$, and

$$\sigma(g_1, g_2) + \sigma(g_1g_2, g_3) = g_1 \cdot \sigma(g_2, g_3) + \sigma(g_1, g_2g_3)$$

for all $g_1, g_2, g_3 \in G$.

Proof The first of these follows from the fact that $s(e_G) = e_E$. For,

$$s(e_Gg) = s(g) = e_Es(g) = s(e_G)s(g)$$

$$\text{and } s(ge_G) = s(g) = s(g)e_E = s(g)s(e_G).$$

The second follows from the associativity condition: for $a_1, a_2, a_3 \in A$ and $g_1, g_2, g_3 \in G$ we require that

$$((a_1, g_1) * (a_2, g_2)) * (a_3, g_3) = (a_1, g_1) * ((a_2, g_2) * (a_3, g_3)).$$

The left hand side yields

$$\begin{aligned} ((a_1, g_1) * (a_2, g_2)) * (a_3, g_3) &= (a_1 + \phi_{g_1}(a_2) + \sigma(g_1, g_2), g_1 g_2) * (a_3, g_3) \\ &= (a_1 + \phi_{g_1}(a_2) + \phi_{g_1 g_2}(a_3) + \sigma(g_1, g_2) + \sigma(g_1 g_2, g_3), g_1 g_2 g_3) \end{aligned}$$

while the right hand side gives

$$\begin{aligned} (a_1, g_1) * ((a_2, g_2) * (a_3, g_3)) &= (a_1, g_1) * (a_2 + \phi_{g_2}(a_3) + \sigma(g_2, g_3), g_2 g_3) \\ &= (a_1 + \phi_{g_1}(a_2 + \phi_{g_2}(a_3) + \sigma(g_2, g_3)) + \sigma(g_1, g_2 g_3), g_1 g_2 g_3). \end{aligned}$$

Setting both of these equal, we get

$$\sigma(g_1, g_2) + \sigma(g_1 g_2, g_3) = g_1 \cdot \sigma(g_2, g_3) + \sigma(g_1, g_2 g_3)$$

as claimed. $\square$

A factor set $\sigma: G \times G \rightarrow A$ thus completely determines the group structure on E for an extension

$$A \xhookrightarrow{i} E \begin{matrix} \xrightarrow{q} \\ \xleftarrow{s} \end{matrix} G$$

of G by A , with a specified section $s: G \rightarrow E$. In order to reconstruct the extension, we therefore need three pieces of information: the group G , the G -module A and the factor set σ . From these, we can build the middle group $E = E[G, A, \sigma]$ and recover the homomorphisms i and q , by setting E to be the group with underlying set $A \times G$ and the multiplication operation defined in (7.23). The inclusion homomorphism i maps any $a \in A$ to the ordered pair (a, e_G) in $E = A \times G$, and the quotient homomorphism q takes any $(a, g) \in E = A \times G$ to the corresponding element $g \in G$.

Example 7.93 Let's use this to construct the cyclic group $\mathbb{Z}_4$ as an extension of $\mathbb{Z}_2$ by $\mathbb{Z}_2$. There is only one $\mathbb{Z}_2$ -action we can define on $\mathbb{Z}_2$, namely the trivial one, so that's the one we'll use.

We want to define a group structure on the set $\mathbb{Z}_2 \times \mathbb{Z}_2$, and to do this we have to choose an appropriate factor set $\sigma: \mathbb{Z}_2 \times \mathbb{Z}_2 \rightarrow \mathbb{Z}_2$. By the first condition in Proposition 7.92, we know that

$$\sigma(0, 0) = \sigma(0, 1) = \sigma(1, 0) = 0,$$

and so the only thing left for us to do is to decide the value of $\sigma(1, 1)$. Setting $\sigma(1, 1) = 0$ yields a trivial factor set, which corresponds to the split extension $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

The other option is to set $\sigma(1, 1) = 1$ (check that this satisfies the second condition in Proposition 7.92). This determines a group structure isomorphic to $\mathbb{Z}_4$ on the set $\mathbb{Z}_2 \times \mathbb{Z}_2$:

	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	(0,1)	(1,0)	(1,1)	(0,0)
(1,0)	(1,0)	(1,1)	(0,0)	(0,1)
(1,1)	(1,1)	(0,0)	(0,1)	(1,0)

This group is cyclic of order 4, with possible generators being either (0,1) or (1,1). The corresponding sections are those for which, respectively, $1 \mapsto (0,1)$ or $1 \mapsto (1,1)$.

Another interesting example relates to the carrying operation in base 10 arithmetic:

Example 7.94 There are two obvious extensions of $\mathbb{Z}_{10}$ by $\mathbb{Z}_{10}$: the direct sum $\mathbb{Z}_{10} \oplus \mathbb{Z}_{10}$ and the cyclic group $\mathbb{Z}_{100}$. We can regard both of these as having the same underlying set $\mathbb{Z}_{10} \times \mathbb{Z}_{10}$ with different addition operations. With the first, the sum $(2,7) + (3,4)$ evaluates to $(5,1)$, while in the second case it evaluates to $(6,1)$ because $27 + 34 = 61$.

In both cases we set $A = G = \mathbb{Z}_{10}$, and consider A to be a trivial $\mathbb{Z}_{10}$ -module. The semidirect product $A \rtimes G$ thus gives the direct sum $\mathbb{Z}_{10} \oplus \mathbb{Z}_{10}$ and the split extension

$$\mathbb{Z}_{10} \xhookrightarrow{i} \mathbb{Z}_{10} \oplus \mathbb{Z}_{10} \xrightarrow{q} \mathbb{Z}_{10}.$$

For the extension

$$\mathbb{Z}_{10} \xhookrightarrow{i} \mathbb{Z}_{100} \xrightarrow{q} \mathbb{Z}_{10},$$

however, we need to introduce a nontrivial factor set that encodes the carrying operation:

$$\sigma(m, n) = \begin{cases} 0 & \text{if } m + n < 10, \\ 1 & \text{if } m + n \geq 10, \end{cases}$$

for all $m, n \in \mathbb{Z}_{10}$. This satisfies the conditions in Proposition 7.92 and is hence a factor set.

What we now need to do is translate the notion of equivalence of extensions in Definition 7.91 into an equivalence relation on factor sets. To do this, we'll investigate how the factor set σ changes if we change the section s . Suppose that

$$A \xhookrightarrow{i} E \xrightarrow{q} G$$

is an extension of a group G by a G -module A . Let s_1 and s_2 be sections of this extension that both realise the specified G -action on A . Then s_1 determines a factor set σ_1 and s_2 yields another factor set σ_2 . For any $g \in G$ there is a unique element $\omega(g) \in A$ such that

$$s_1(g) = i(\omega(g))s_2(g).$$

This determines a function $\omega: G \rightarrow A$ with $\omega(e_G) = 0$. Then, for any $g, h \in G$,

$$s_1(g)s_1(h) = i(\omega(g))s_2(g)i(\omega(h))s_2(h),$$

and also

$$\begin{aligned} s_1(g)s_1(h) &= s_1(gh)i(\sigma_1(g, h)) \\ &= s_2(gh)i(\omega(gh))i(\sigma_1(g, h)) \\ &= s_2(gh)i(\omega(gh) + \sigma_1(g, h)). \end{aligned}$$

Hence, in terms of the group structure on $A \times G$,

$$(\omega(g), g) * (\omega(h), h) = (\omega(gh) + \sigma_1(g, h), gh).$$

The left hand side of this gives

$$(\omega(g), g) * (\omega(h), h) = (\omega(g) + g \cdot \omega(h) + \sigma_2(g, h), gh).$$

Equating these, we get

$$\omega(gh) + \sigma_1(g, h) = \omega(g) + g \cdot \omega(h) + \sigma_2(g, h)$$

and hence

$$\sigma_1(g, h) - \sigma_2(g, h) = \omega(g) + g \cdot \omega(h) - \omega(gh). \quad (7.24)$$

What this means is that different sections of the same group extension yield factor sets that differ by an expression of the form given in (7.24), which depends on a special sort of function:

Definition 7.95 Let G be a group, and A a G -module. A **coboundary** or **inner factor set** is a function $\tau: G \times G \rightarrow A$ where

$$\tau(g, h) = \omega(g) - \omega(gh) + g \cdot \omega(h)$$

for all $g, h \in G$, and some function $\omega: G \rightarrow A$ satisfying $\omega(e_G) = 0$.

Two factor sets σ_1 and σ_2 are **equivalent** or **cohomologous** if their difference is an inner factor set τ .

It is relatively straightforward to verify that an inner factor set satisfies the criteria in Proposition 7.92 and is hence a factor set.

Now, given a group G and G -module A , we can form the set $\text{FS}(G, A)$ of factor sets $\sigma: G \times G \rightarrow A$, and the subset $\text{IFS}(G, A)$ of inner factor sets. More interestingly, these sets are closed under a fairly straightforward addition operation and thereby form abelian groups:

Proposition 7.96 Let G be a group and A a G -module. Suppose that $\sigma_1, \sigma_2 \in \text{FS}(G, A)$ are factor sets. Then the function $\sigma_1 + \sigma_2$ given by $(\sigma_1 + \sigma_2)(g, h) = \sigma_1(g, h) + \sigma_2(g, h)$ is also a factor set. Furthermore, if $\tau_1, \tau_2 \in \text{IFS}(G, A)$ then $\tau_1 + \tau_2 \in \text{IFS}(G, A)$ too.

Proof To see that $\text{FS}(G, A)$ and $\text{IFS}(G, A)$ are closed under this pointwise addition operation, we must verify that $\sigma_1 + \sigma_2$ satisfies the conditions in Proposition 7.92, and that $\tau_1 + \tau_2$ satisfies Definition 7.95.

Both of these tasks are entirely routine, and we'll omit the details here, although the interested reader is encouraged to check them. $\square$

Let's summarise where we are at the moment. For any extension

$$A \xhookrightarrow{i} E \xrightarrow{q} G$$

of a group G by a G -module A , we can choose one or more **sections** $s: G \rightarrow E$ that **realise** the specified G -module structure on A . Each of these sections determines a **factor set** $\sigma: G \times G \rightarrow A$, and if we throw away s and just keep G , A and σ , we can reconstruct E in a canonical way. If s is a homomorphism then $\sigma = 0$ and $E \cong A \rtimes G$, and in this case we have a **split extension**. Different sections of the same extension yield factor sets whose difference is an **inner factor set**. Furthermore, the sum of two factor sets is also a factor set, and the sum of two inner factor sets is also an inner factor set; this enables us to define groups $\text{FS}(G, A)$ and $\text{IFS}(G, A)$. The quotient $\text{Ext}(G, A) = \text{FS}(G, A) / \text{IFS}(G, A)$ consists of equivalence classes of factor sets that differ only by an inner factor set.³¹ We claim that this group $\text{Ext}(G, A)$ *classifies* extensions of G by A : there is a bijective correspondence between elements of $\text{Ext}(G, A)$ and equivalence classes of extensions of G by A in the sense of Definition 7.91. We're almost there: all we need to do is to show that equivalent extensions correspond to **equivalent** (or **cohomologous**) factor sets, and vice versa.

³¹ Some books, in particular those on homological algebra, call factor sets **co-cycles** and inner factor sets **coboundaries**; they form groups $Z^2(G; A)$ and $B^2(G; A)$, respectively, and their quotient $H^2(G; A)$ is the second in a sequence of **cohomology groups**, each of which contains important structural information. Homological algebra is an interesting and powerful branch of mathematics with strong links to algebraic topology, but a full discussion is beyond the scope of this book.

Proposition 7.97 *Two extensions*

$$A \xhookrightarrow{i_1} E_1 \xrightarrow{q_1} G \quad \text{and} \quad A \xhookrightarrow{i_2} E_2 \xrightarrow{q_2} G$$

of a group G by a G -module A are equivalent if and only if they have equivalent factor sets.

Proof Suppose that the extensions are equivalent; that is, there exists an isomorphism $\phi: E_1 \rightarrow E_2$ such that $\phi \circ i_1 = i_2$ and $q_1 = q_2 \circ \phi$. Choose a section $s_1: G \rightarrow E_1$ that realises the G -action on A . Then s_1 yields a factor set σ for E_1 .

For any $g \in G$, the isomorphism ϕ maps the coset $i_1(A)s_1(g)$ to the coset

$$\phi(i_1(A)s_1(g)) = \phi(i_1(A))\phi(s_1(g)) = i_2(A)\phi(s_1(g))$$

in E_2 . Define a section $s_2: G \rightarrow E_2$ by $s_2 = \phi \circ s_1$. We need to check that this section also realises the given G -module structure on A . The required G -action is realised by s_1 , so for any $g \in G$ and $a \in A$ we have

$$i_1(g \cdot a) = s_1(g)i_1(a)s_1(g)^{-1}.$$

Applying ϕ to both sides of this equation, we get

$$i_2(g \cdot a) = \phi(i_1(g \cdot a)) = \\ \phi(s_1(g))\phi(i_1(a))\phi(s_1(g))^{-1} = s_2(g)i_2(a)s_2(g)^{-1}$$

and so s_2 does indeed realise the correct G -action.

The factor set σ is determined by the section s_1 as follows:

$$i_1(\sigma(g, h)) = s_1(g)s_1(h)s_1(gh)^{-1}$$

for all $g, h \in G$. Applying ϕ to both sides, we get

$$i_2(\sigma(g, h)) = s_2(g)s_2(h)s_2(gh)^{-1}$$

so the factor set determined by the section s_2 is the same as that determined by s_1 . If τ is some other factor set for E_2 determined by a different section $t: G \rightarrow E_2$, then by (7.24) τ and σ are cohomologous. Conversely, suppose that E_1 and E_2 are extensions of G by A relative to equivalent factor sets σ and τ respectively. Then there exists a function $\omega: G \rightarrow A$ such that

$$\sigma(g, h) - \tau(g, h) = \omega(g) + g \cdot \omega(h) - \omega(gh)$$

for all $g, h \in G$.

Let $s_1: G \rightarrow E_1$ and $s_2: G \rightarrow E_2$ be sections that realise the G -module structure on A and yield the factor sets σ and τ respectively. Then every element in E_1 can be written uniquely as $i_1(a)s_1(g)$ for some $a \in A$ and $g \in G$, and every element in E_2 can be similarly written in the form $i_2(a)s_2(g)$. Recall that the multiplication operation in E_1 is given by

$$i_1(a)s_1(g) * i_1(b)s_1(h) = i_1(a + g \cdot b + \sigma(g, h))s_1(gh)$$

and that in E_2 is given by

$$i_2(a)s_2(g) * i_2(b)s_2(h) = i_2(a + g \cdot b + \tau(g, h))s_2(gh).$$

Now define $\phi: E_1 \rightarrow E_2$ by $\phi(i_1(a)s_2(g)) = i_2(a + \omega(g))s_2(g)$. This is clearly a bijection. Furthermore

$$\begin{aligned} \phi(i_1(a)s_1(g) * i_1(b)s_1(h)) &= \phi(i_1(a + g \cdot b + \sigma(g, h))s_1(gh)) \\ &= i_2(a + g \cdot b + \sigma(g, h) + \omega(gh))s_2(gh) \end{aligned}$$

and

$$\begin{aligned} \phi(i_1(a)s_1(g)) * \phi(i_1(b)s_1(h)) &= i_2(a + \omega(g))s_2(g) * i_2(b + \omega(h))s_2(h) \\ &= i_2(a + \omega(g) + g \cdot b + g \cdot \omega(h) + \tau(g, h))s_2(gh). \end{aligned}$$

Since σ and τ are equivalent,

$$\sigma(g, h) + \omega(gh) = \omega(g) + g \cdot \omega(h) + \tau(g, h)$$

and hence the two expressions above are equal, which means that ϕ is an isomorphism.



Oberwolfach Photo Collection / Konrad Jacobs

Issai Schur (1875–1941)

Next we need to show that the diagram in Definition 7.91 commutes. For any $a \in A$ and $g \in G$,

$$\phi(i_1(a)) = \phi(i_1(a)s_1(e_G)) = i_2(a + \omega(e_G))s_2(e_G) = i_2(a)$$

and

$$q_2(\phi(i_1(a)s_1(g))) = q_2(i_2(a + \omega(g))s_1(g)) = g = q_1(i_1(a)s_1(g))$$

as required. Hence E_1 and E_2 are equivalent. $\square$

So, the group $\text{Ext}(G, A)$ classifies extensions of a group G by a G -module A . The zero element in $\text{Ext}(G, A)$ corresponds to the (equivalence class of) the split extension

$$A \xhookrightarrow{i} A \rtimes G \xrightarrow{q} G$$

and so if $\text{Ext}(G, A)$ is trivial, this is, up to equivalence, the only extension of G by A .

The next result is due originally to the German algebraist Issai Schur (1875–1941) and was later extended to the nonabelian case by Hans Zassenhaus (1912–1991).

Theorem 7.98 (Schur–Zassenhaus Theorem) *Let G be a group of order m and A a G -module of order n . If $\gcd(m, n) = 1$ then $\text{Ext}(G, A)$ is trivial, and the only extension of G by A is the semidirect product $A \rtimes G$.*

Proof Suppose that σ is a factor set in $\text{FS}(G, A)$; recall that this is a function $\sigma: G \times G \rightarrow A$, and hence its values lie in A . By Proposition 2.34 these values must all have orders that divide $|A| = n$, and so $n\sigma(g, h) = 0$ for all $g, h \in G$. Therefore $n\sigma = 0$ in $\text{FS}(G, A)$, and $n\sigma$ also lies in the subgroup $\text{IFS}(G, A)$ of inner factor sets.

Now we want to show that $m\sigma$ is also an inner factor set. To do this, we define a function $\omega: G \rightarrow A$ by

$$\omega(g) = \sum_{h \in G} \sigma(g, h).$$

In particular,

$$\omega(e_G) = \sum_{h \in G} \sigma(e_G, h) = \sum_{h \in G} 0 = 0.$$

Since σ is a factor set, we have

$$\sigma(g, h) = \sigma(g, hk) + g \cdot \sigma(h, k) - \sigma(gh, k)$$

for all $g, h, k \in G$. Sum both sides of this equation as k ranges over the whole of G :

$$\sum_{k \in G} \sigma(g, h) = \sum_{k \in G} \sigma(g, hk) + g \cdot \sum_{k \in G} \sigma(h, k) - \sum_{k \in G} \sigma(gh, k).$$

This gives

$$m\sigma(g, h) = \omega(g) + g \cdot \omega(h) - \omega(gh)$$

and hence $m\sigma$ is an inner factor set as well.

Since m and n are coprime, we can find integers p and q such that $pm + qn = 1$. Then

$$\sigma = (pm + qn)\sigma = p(m\sigma) + q(n\sigma)$$

lies in $\text{IFS}(G, A)$. This means that any factor set $\sigma \in \text{FS}(G, A)$ is inner, so $\text{FS}(G, A) = \text{IFS}(G, A)$ and hence the quotient $\text{Ext}(G, A) = \text{FS}(G, A) / \text{IFS}(G, A)$ is trivial. $\square$

We will now look at a few specific types of group extensions.

First, suppose that

$$A \xhookrightarrow{i} A \rtimes G \xrightarrow{q} G$$

is an extension of a group G by a G -module A , and that $s: G \rightarrow E$ is a section that realises the given G -module structure on A . Then by rearranging (7.20) we get the commutation relation (7.21), reproduced here:

$$i(g \cdot a)s(g) = s(g)i(a) \quad (7.25)$$

What this says is that for each $i(a)$ to commute with all the $s(g)$ (and, for that matter, every other element of E) we require $g \cdot a = a$; that is, for the G -module structure on A to be trivial. This means that every element of the form $i(a)$ commutes with every element of E , and therefore that the image $i(A)$ of A lies in the centre $Z(E)$ of E . Such extensions are called **central extensions**, and the triviality of the G -action on A is both a necessary and sufficient condition for an extension to be central.

Example 7.99 The quaternion group Q_8 is a central extension of the Klein group $V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$ by the cyclic group $\mathbb{Z}_2$. The latter embeds into Q_8 as the subgroup $\{\pm E\}$ and one possible quotient map takes $i(\mathbb{Z}_2) = \{\pm E\}$ to the identity $e \in V_4$, the coset $\{\pm I\}$ to $a \in V_4$, the coset $\{\pm J\}$ to $b \in V_4$, and the coset $\{\pm K\}$ to $c \in V_4$.

A possible section $s: V_4 \rightarrow Q_8$ is given by

$$s: \begin{cases} e \mapsto E, & a \mapsto I, \\ b \mapsto J, & c \mapsto K \end{cases}$$

and the corresponding factor set is:

σ	e	a	b	c
e	0	0	0	0
a	0	1	0	1
b	0	1	1	0
c	0	0	1	1

Another interesting case is the one where the group G is cyclic. These are called **cyclic extensions**, and the following discussion will be useful when we classify groups of order 16 in Section 7.A.

If we have a group G with a normal subgroup $H \trianglelefteq G$, such that $G/H \cong \mathbb{Z}_n$ then G is obviously a cyclic extension of $\mathbb{Z}_n$ by some (possibly nonabelian) group H . Choose some element $g \in G \setminus H$ such that the coset gH generates the quotient G/H . Let $v = g^n$. Then $vH = g^nH = (gH)^n = H$ in G/H , and so $v \in H$. As a normal subgroup, H is closed under conjugation by any element of G , so let $\tau \in \text{Aut}(H)$ be the inner automorphism $\tau: h \mapsto ghg^{-1}$. Furthermore,

$$\tau(v) = gv g^{-1} = gg^n g^{-1} = g^n = v$$

so τ fixes this distinguished element v . Also, for any $h \in H$,

$$\tau^n(h) = g^n h g^{-n} = v h v^{-1}$$

so τ^n is conjugation by v .

We can now discard G and g , and just keep v . We'll now see that the data (H, n, τ, v) is all we need to reconstruct and uniquely determine (up to equivalence) the extension

$$H \xhookrightarrow{i} G \xrightarrow{q} \mathbb{Z}_n.$$

Definition 7.100 A **cyclic extension type** for a group H is a quadruple (H, n, τ, v) where $n \in \mathbb{N}$, $\tau \in \text{Aut}(H)$ and $v \in H$ such that $\tau(v) = v$ and τ^n is conjugation by v .

Every cyclic extension type determines some group G , and all cyclic extensions arise in this way:

Theorem 7.101 (Cyclic Extension Theorem) *Given a cyclic extension type (H, n, τ, v) there is some group G with $H \trianglelefteq G$ and $G/H \cong \mathbb{Z}_n$. Furthermore, all extensions of $\mathbb{Z}_n$ by H are determined by some cyclic extension type.*

Proof Suppose that $\mathbb{Z}_n = \langle t : t^n = 1 \rangle$. Let G be the set of ordered pairs (h, t^i) where $h \in H$ and $0 \leq i < n$. We want to define a group multiplication structure on this set, which we do as follows:

$$(h, t^i) * (k, t^j) = (h\tau^i(k)\sigma(i, j), t^{i+j})$$

with (possibly nonabelian) factor set

$$\sigma(i, j) = \begin{cases} e_H & \text{if } i + j < n, \\ v & \text{if } i + j \geq n. \end{cases}$$

This operation is associative, the identity element is the pair (e_H, t^0) , and the inverse of (h, t^i) is $(\tau^{-i}((vh)^{-1}))$. These can all be checked by a straightforward but involved calculation which we will omit.

This group is an extension of $\mathbb{Z}_n$ by H , since the quotient map $q: G \rightarrow G/H$ maps (h, t^i) to $t^i H \in G/H$, and $H = \{(h, t^0) : h \in H\}$ is normal in G . To see the latter, observe that

$$\begin{aligned} (k, t^i) * (h, t^0) * (k, t)^{-1} &= (k, t^i) * (h, t^0) * (\tau^{-1}((vk)^{-1}), t^{n-i}) \\ &= (k\tau^i(h), t^i) * (\tau^{-i}((vk)^{-1}), t^{n-i}) \\ &= (k\tau^i(h)\tau^i(\tau^{-i}((vk)^{-1})), t^{i+n-i-n}) \\ &= (k\tau^i(h)(vk)^{-1}, t^0) \in H \times \{t^0\} \subseteq G \end{aligned}$$

and so H is closed under conjugation and hence normal in G .

Conversely, suppose that some group G has a normal subgroup H and that $G/H \cong \mathbb{Z}_n$ for some n . Then by the discussion earlier, we can choose an element $g \in G \setminus H$ such that $G/H = \langle gH \rangle$, and from this obtain an automorphism $\tau \in \text{Aut}(H)$, defined by $\tau: h \mapsto ghg^{-1}$, and an element $v = g^n \in H$ with the required properties. $\square$

So, given a cyclic group $\mathbb{Z}_n$ and a group H of order m , we can construct a list of groups of order mn by listing all possible cyclic extension types (H, n, τ, v) . This list might contain duplicates, and it might not contain all possible groups of order mn , but it's a start. To solve the duplication problem we need an appropriate notion of equivalence for cyclic extension types.

Definition 7.102 Let (H, n, τ, v) and (K, m, σ, w) be cyclic extension types. They are **equivalent** if $m = n$ and there exists an isomorphism $\phi: H \rightarrow K$ such that $\phi \circ \tau = \sigma \circ \phi$ and $w = \phi(v)$.

This concept of equivalence is exactly the one we want:

Proposition 7.103 Equivalent cyclic extension types yield isomorphic groups.

Proof Let (H_1, n, τ_1, v_1) and (H_2, n, τ_2, v_2) be equivalent cyclic extension types. Then there exists an isomorphism $\phi: H_1 \rightarrow H_2$ satisfying the conditions in Definition 7.102. Let G_1 and G_2 be groups realising, respectively, (H_1, n, τ_1, v_1) and (H_2, n, τ_2, v_2) . Then there are elements $g_1 \in G_1 \setminus H_1$ and $g_2 \in G_2 \setminus H_2$ such that τ_1 is conjugation by g_1 and τ_2 is conjugation by g_2 , and $v_1 = g_1^n$ and $v_2 = g_2^n$.

Let $f: G_1 \rightarrow G_2$ such that $hg_1^i \mapsto \phi(h)g_2^i$; this is well defined and bijective. We must show that it is also a homomorphism. Suppose that $i + j < n$. Then

$$\begin{aligned} f((hg_1^i) * (kg_1^j)) &= f(h\tau_1(k)g_1^{i+j}) = \phi(h\tau_1(k))g_2^{i+j} \\ &= \phi(h)\tau_2(\phi(k))g_2^{i+j} = f(hg_1^i) * f(kg_1^j). \end{aligned}$$

If $i + j \geq n$ then

$$f((hg_1^i) * (kg_1^j)) = f(h\tau_1(k)v_1g_1^{i+j}) = \phi(h\tau_1(k)v_1)g_2^{i+j}$$

$$= \phi(h)\tau_2(\phi(k))v_2g_2^{i+j} = f(hg_1^i) * f(kg_1^j).$$

Hence $f: G_1 \rightarrow G_2$ is an isomorphism, as claimed. $\square$

Note that the converse doesn't necessarily hold: there may be inequivalent cyclic extension types that also realise isomorphic groups.

Proposition 7.104 *Let (H, n, τ, v) and (H, n, σ, v) be two cyclic extension types with τ and σ conjugate in $\text{Aut}(H)$. Then any group that realises (H, n, τ, v) is isomorphic to any group that realises (H, n, σ, v)*

Proof If τ and σ are conjugate in $\text{Aut}(H)$ then there exists an automorphism $\phi \in \text{Aut}(H)$ such that $\sigma = \phi \circ \tau \circ \phi^{-1}$. Then $\phi: H \rightarrow H$ is an equivalence of cyclic extension types, since $\phi \circ \tau = \sigma \circ \phi$, and since $\phi(v) = v$ it follows that (H, n, τ, v) and (H, n, σ, v) are equivalent. Hence, by Proposition 7.103, they are realised by isomorphic groups. $\square$

We end this part of the discussion with a short example illustrating the use of cyclic extension types for classification of groups of a given order.

Example 7.105 We can use these techniques to classify the groups of order 6 again. By Cauchy's Theorem, a group G of order 6 must have a subgroup $H \cong \mathbb{Z}_3$, which has index 2 and is therefore normal by Proposition 3.12. The quotient $G/H \cong \mathbb{Z}_2$ and so the problem reduces to studying cyclic extension types of the form $(H, 2, \tau, v)$.

There are two automorphisms of $\mathbb{Z}_3$: the identity, which we will denote τ_1 here, and the map τ_2 that takes every element of H to its inverse. From these we obtain four cyclic extension types:

$$(\mathbb{Z}_3, 2, \tau_1, 1), \quad (\mathbb{Z}_3, 2, \tau_1, h), \quad (\mathbb{Z}_3, 2, \tau_1, h^2) \quad \text{and} \quad (\mathbb{Z}_3, 2, \tau_2, 1).$$

The first three of these are realised by the direct sum $\mathbb{Z}_3 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_6$, while the last is realised by the dihedral group D_3 .

We'll use cyclic extension types in the next section, when we classify groups of order 16. Finally, we'll mention a couple of classes of groups that can be described in terms of extensions.

Definition 7.106 A **metabelian** group is a group that can be constructed as an extension of an abelian group by an abelian group; that is, a group G that has an abelian normal subgroup N such that the quotient G/N is also abelian.

A **metacyclic group** is a group that is isomorphic to a cyclic extension by a cyclic group: a group G with a cyclic normal subgroup N such that G/N is also cyclic.

Metacyclic groups are metabelian, and include the cyclic groups themselves, the dicyclic groups Dic_n and the dihedral groups D_n . As

remarked earlier, metabelian groups are exactly those with derived length at most two.

7.A Classification of small finite groups

IN THIS SECTION WE WILL CLASSIFY, up to isomorphism, groups of order less than 32. The reason we'll stop at order 31 is that the classification is relatively manageable up to that point. The classification of groups of order $n = 2^k$ is, in general, quite complicated and involves the consideration of a number of cases. As can be seen from Table 7.3, the only really involved cases are $n = 16 = 2^4$, with 14 non-isomorphic groups, and $n = 24 = 3 \times 2^3$, with 15. Everything else is comparatively straightforward; however there are 51 groups of order 32, which is why we'll stop just before then.

Before we start classifying, we'll introduce a few results that will prove useful. The first relates to groups that decompose as a semidirect product of two cyclic groups.

Proposition 7.107 Suppose that a finite group G decomposes as an internal semidirect product HK , where $H = \langle x : x^m = 1 \rangle \cong \mathbb{Z}_m$ is normal in G , and $K = \langle y : y^n = 1 \rangle \cong \mathbb{Z}_n$. Then the possible actions of K on H are determined by $yxy^{-1} = x^i$ where $0 < i < m$ and $i^n \equiv 1 \pmod{m}$.

Proof The action of K on H is determined entirely by the behaviour of the generators x and y , specifically by the value of $yxy^{-1} \in H$. Since H is cyclic, this must be a nontrivial power of the generator x , so $yxy^{-1} = x^i$ with $0 < i < m$. Furthermore, since K is a finite cyclic group of order n , we have $y^n = 1$, so $x = y^nxy^{-n} = x^{i^n}$ and therefore $i^n \equiv 1 \pmod{m}$. $\square$

The next result, which we will use particularly in the classification of groups of order 16 and 27, relates to factoring a group by its centre.

Proposition 7.108 If $G/Z(G)$ is cyclic then G is abelian.

Proof Let $Z = Z(G)$ and suppose that G/Z is cyclic. Then there exists some $g \in G$ such that $G/Z = \langle gZ \rangle$, and hence every coset of Z in G is of the form $(gZ)^k = g^kZ$ for some $k \in \mathbb{Z}$.

Consider $x, y \in G$ with $x \in g^mZ$ and $y \in g^nZ$. Then $x = g^ma$ and $y = g^nb$ for some $a, b \in Z$. Therefore

$$xy = g^ma g^nb = g^m g^nb a = g^{m+n}ba = g^n g^mba = g^nb g^ma = yx$$

and hence G is abelian. $\square$

I tried to make out the names of plants, and collected all sorts of things, shells, seals, franks, coins and minerals. The passion for collecting, which leads a man to be a systematic naturalist, a virtuoso, or a miser, was very strong in me, and was clearly innate, as none of my sisters or brother ever had this taste.

— Charles Darwin (1809–1882), in: Francis Darwin, *The Life and Letters of Charles Darwin* (1887) I 27–28

n	total	abelian	nonabelian
1	1	1	0
2	1	1	0
3	1	1	0
4	2	2	0
5	1	1	0
6	2	1	1
7	1	1	0
8	5	3	2
9	2	2	0
10	2	1	1
11	1	1	0
12	5	2	3
13	1	1	0
14	2	1	1
15	1	1	0
16	14	5	9
17	1	1	0
18	5	2	3
19	1	1	0
20	5	2	3
21	2	1	1
22	2	1	1
23	1	1	0
24	15	3	12
25	2	2	0
26	2	1	1
27	5	3	2
28	4	2	2
29	1	1	0
30	4	1	3
31	1	1	0

Table 7.3: Groups of orders $n = 1$ –31

The trivial group

There is a single group with one element: the trivial group $\{e\}$.

Groups of prime order: 2, 3, 5, 7, 11, 13, 17, 19, 23, 29 and 31

By Proposition 2.35, any group G of prime order $|G| = p$ is isomorphic to the cyclic group $\mathbb{Z}_p$. These groups are abelian, simple and soluble.

Groups of twice prime order: 4, 6, 10, 14, 22 and 26

The groups of order 4 were classified in Proposition 2.38 and those of order 6 in Proposition 2.39. More generally, by Proposition 5.29, there are two groups of order $2p$ where p is prime, listed in Table 7.4.

The cyclic group $\mathbb{Z}_{2p}$ is abelian, and therefore soluble by Proposition 7.42. The dihedral group D_p (in the case $p = 2$ this is the Klein group V_4) is soluble by Proposition 7.40 because it is an extension of two finite abelian (and hence soluble) groups. More precisely, D_p has a soluble normal subgroup $R_p \cong \mathbb{Z}_p$ and soluble quotient $D_p/R_p \cong \mathbb{Z}_2$.

(i)	The cyclic group $\mathbb{Z}_{2p}$ and
(ii)	the dihedral group D_p .

Table 7.4: Groups of order $2p$

Groups of prime square order: 4, 9 and 25

By Proposition 7.20, if p is prime, then there are two groups of order p^2 , listed in Table 7.5, both of which are abelian. These are soluble by Propositions 7.42 and 7.43.

(i)	The cyclic group $\mathbb{Z}_{p^2}$ and
(ii)	$\mathbb{Z}_p \oplus \mathbb{Z}_p$.

Table 7.5: Groups of order p^2

Groups of order 8

The groups of order 8 were classified in Proposition 2.40, and there are five of them, listed in Table 7.6. Of these, (i)–(iii) are abelian, while (iv) and (v) are nonabelian. They are all soluble by Proposition 7.43.

(i)	the cyclic group $\mathbb{Z}_8$,
(ii)	the abelian groups $\mathbb{Z}_2 \oplus \mathbb{Z}_4$ and
(iii)	$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$,
(iv)	the dihedral group D_4 and
(v)	the quaternion group Q_8 , which is isomorphic to the dicyclic group Dic_2 .

Table 7.6: Groups of order 8

Groups of order 12

This is the first new case that can't be immediately resolved by existing classification results. Groups of order 12 that we've met already include the cyclic group $\mathbb{Z}_{12} \cong \mathbb{Z}_3 \oplus \mathbb{Z}_4$, the abelian group $\mathbb{Z}_2 \oplus \mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3$, the dihedral group D_6 , the alternating group A_4 and the dicyclic group Dic_3 . We now prove that these five are the only ones.

Let G be a group of order $|G| = 12$. First we apply Sylow's Third Theorem³² to calculate the possible numbers n_p of Sylow p -subgroups.

³² Theorem 7.13, page 240.

Thus $n_2 \equiv 1 \pmod{2}$ and $n_2|3$, which means that either $n_2 = 1$ or 3 . Similarly, $n_3 \equiv 1 \pmod{3}$ and $n_3|4$, so either $n_3 = 1$ or 4 .

Case 1 ($n_3 = 1$) Let H be the (unique) Sylow 3-subgroup; this must be isomorphic to $\mathbb{Z}_3$. Furthermore, $H \triangleleft G$ by Corollary 7.12. Now let K be a Sylow 2-subgroup: this has order 4 and must therefore be isomorphic to either $\mathbb{Z}_4$ or $\mathbb{Z}_2$.

By Proposition 2.34, the order of any element of $H \cap K$ must divide both $|H| = 3$ and $|K| = 4$, which are coprime, so $H \cap K = \{e\}$. Then $G = HK$ must be a semidirect product of H and K . Let $H = \langle x : x^3 = 1 \rangle$ and consider the two possibilities for K separately:

Case 1a ($K \cong \mathbb{Z}_4$) Let $K = \langle y : y^4 = 1 \rangle$. By Proposition 7.107 the K -action is determined by the value of $xyx^{-1} = x^i$, where $0 < i < 3$ and $i^4 \equiv 1 \pmod{3}$. There are two ways this can go: either $xyx^{-1} = x$ or $xyx^{-1} = x^2 = x^{-1}$.

The first of these gives the presentation

$$G = \langle x, y : x^3 = y^4 = 1, xy = yx \rangle \cong \mathbb{Z}_4 \oplus \mathbb{Z}_3 \cong \mathbb{Z}_{12}.$$

The other possibility is that $xyx^{-1} = x^2$, which gives the presentation

$$G = \mathbb{Z}_3 \rtimes \mathbb{Z}_4 = \langle x, y : x^3 = y^4 = 1, x^2y = yx \rangle.$$

We claim that this group is isomorphic to the dicyclic group

$$\text{Dic}_3 = \langle a, b : a^6 = e, a^3 = b^2, ab = ba^{-1} \rangle.$$

To see this, we note first that

$$y^2x = y(yx) = y(x^2y) = (yx)xy = x^2(yx)y = x^4y^2 = xy^2$$

and then that $(y^2x)^2 = (y^2x)y^2x = xy^4x = x^2$. Since x^2 has order 3, it follows that y^2x has order 6. Now let $a = y^2x$ and $b = y$. Then $b^2a = y^4x = x$, so a and b generate all of the group $G = \mathbb{Z}_3 \rtimes \mathbb{Z}_4$. Furthermore,

$$a^3 = (y^2x)^3 = x^2y^2x = x^3y^2 = y^2 = b^2$$

$$\text{and } aba = y^2xy^3x = xy^5x = xyx = x^3y = y = b.$$

So $G \cong \text{Dic}_3$.

Case 1b ($K \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$) Let $K = \langle y, z : y^2 = z^2 = 1, yz = zy \rangle$. The semidirect product $H \rtimes K$ is determined by the conjugation action of b and c on a . There are two possibilities for each:

$$\begin{array}{ll} yxy^{-1} = x & \text{or} \quad yxy^{-1} = x^2 \\ \text{and} \quad zxz^{-1} = x & \text{or} \quad zxz^{-1} = x^2. \end{array}$$

If both y and z act trivially on x , we have

$$G = \langle x, y, z : x^3 = y^2 = z^2 = 1, xy = yx, xz = zx, yz = zy \rangle.$$

This is isomorphic to the direct sum $\mathbb{Z}_3 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_6$.

If $xyx^{-1} = x^2$ and $zxx^{-1} = x$ then we have

$$G = \langle x, y, z : x^3 = y^2 = z^2 = 1, x^2y = yx, xz = zx, yz = zy \rangle. \quad (7.26)$$

This is isomorphic to the case $xyx^{-1} = x$ and $zxx^{-1} = x^2$ by swapping the generators y and z .

The other possibility is that both y and z act nontrivially on x , in which case we get

$$G = \langle x, y, z : x^3 = y^2 = z^2 = 1, x^2y = yx, x^2z = zx, yz = zy \rangle.$$

But swapping y and yz yields a presentation isomorphic to (7.26), so as long as either or both of y and z act nontrivially on x we get isomorphic groups.

We now claim that this group is isomorphic to the dihedral group D_6 . Setting $a = xy$ and $b = yz$, it follows that $a^2 = (xy)^2 = xyxy = x^2y^2 = x^2$. And x^2 has order 3, so $a = xy$ must have order 6. Moreover, $ba = yzxy = yxzy = yxyz = (x^2y)(yz) = a^{-1}b$. The presentation (7.26) can thus be shown to be isomorphic to

$$\langle a, b : a^6 = b^2 = 1, ba = a^{-1}b \rangle$$

which by Proposition 5.28 is isomorphic to D_6 . Also (7.26) can be rewritten as

$$\langle x, y : x^3 = y^2 = 1, x^2y = yx \rangle \times \langle z : z^2 = 1 \rangle,$$

which is a presentation for the direct product $D_3 \times \mathbb{Z}_2$ by Proposition 5.22, and hence $D_6 \cong D_3 \times \mathbb{Z}_2$.

Case 2 ($n_3 = 4$) Suppose instead that G contains four Sylow 3-subgroups, which we can denote H_1, H_2, H_3 and H_4 . By Sylow's Second Theorem they are all conjugate to each other, so we can let G act on $X = \text{Syl}_3(G) = \{H_1, H_2, H_3, H_4\}$ by conjugation. The Orbit-Stabiliser Theorem³³ says that

$$4 = |X| = |\text{Orb}_G(H_i)| = |G|/|\text{Stab}_G(H_i)| = |G|/|N_G(H_i)|$$

and hence $|N_G(H_i)| = 3$, so the normaliser $N_G(H_i) = H_i$ itself. (Thus each of the Sylow 3-subgroups are self-normalising.)

Now suppose that $H_1 = \langle x : x^3 \rangle$. In general, $H_i \cap H_j = \{e\}$ for all $i \neq j$, and so $a \notin N_G(H_i)$ for $2 \leq i \leq 4$. That is, $xH_1x^{-1} = H_1$ but $xH_ix^{-1} \neq H_i$ remained for $2 \leq i \leq 4$. The action of x on X then fixes H_1 and cyclically permutes the other three groups H_2, H_3 and H_4 : it corresponds to the permutation $(H_2 H_3 H_4) \in \text{Sym}(X)$.

Similarly, let $H_4 = \langle y : y^3 \rangle$. Then by a similar argument, the generator y corresponds to the permutation $(H_1 H_2 H_3) \in \text{Sym}(X)$. Therefore G contains $(H_2 H_3 H_4)$ and $(H_1 H_2 H_3)$, which between them generate the alternating group $\text{Alt}(X) \cong A_4$. And since $|A_4| = 12 = |G|$ it follows that $G \cong A_4$.

³³ Theorem 6.25, page 217.

Up to isomorphism, then, there are five groups of order 12, listed in Table 7.7. Of these, the first two are abelian and the other three aren't. All of them are soluble: the first two by Proposition 7.42 because they're abelian, the next two because they are extensions of abelian (and hence soluble) groups by Proposition 7.40, and the last by Proposition 7.38 because it's a subgroup of S_4 , whose derived series (7.18) terminates in the trivial subgroup and is therefore soluble.

(i)	the cyclic group $\mathbb{Z}_{12} \cong \mathbb{Z}_4 \oplus \mathbb{Z}_3$,
(ii)	the abelian group $\mathbb{Z}_2 \oplus \mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3$,
(iii)	the dihedral group D_6 ,
(iv)	the dicyclic group Dic_3 , and
(v)	the alternating group A_4 .

Table 7.7: Groups of order 12

Groups of order 15

By Example 7.14, and more generally by Proposition 7.15, there is only one group of order 15: the cyclic group $\mathbb{Z}_{15} \cong \mathbb{Z}_3 \oplus \mathbb{Z}_5$. This group is abelian, and hence soluble by Proposition 7.42.

Groups of order 16

This is the first of the two complicated cases we will consider (the other being 24). The following exposition is based heavily on the very readable article by Marcel Wild,³⁴ and makes use of the theory of cyclic extensions developed in the previous section.

First we prove the following result:

Proposition 7.109 *Let G be a group of order 16. If G is not isomorphic to the direct sum $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, then G contains a normal subgroup isomorphic either to $\mathbb{Z}_8$ or $\mathbb{Z}_2 \oplus \mathbb{Z}_4$.*

Proof Suppose that $G \not\cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$. If G has an element g of order 8 then the cyclic subgroup $\langle g \rangle \cong \mathbb{Z}_8$; it has index 2 in G and is thus normal by Proposition 3.12.

If G has no element of order 8, then it must have at least some elements of order 4, otherwise G would be isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$. Since G is a p -group (for $p = 2$), by Proposition 7.19 the centre $Z(G)$ must be nontrivial, by Lagrange's Theorem it must have even order, and by Cauchy's Theorem there is an element $z \in Z(G)$ with order 2. Let $H = \langle z \rangle \cong \mathbb{Z}_2$ be the subgroup of G generated by this element. This subgroup is normal in G because it lies in the centre $Z(G)$.

Case 1 Suppose there exists an element $x \in G$ of order 4 such that $x^2 \neq z$. Then the intersection $\langle x \rangle \cap H = \{e\}$ is trivial, and hence $\langle x, z \rangle = \mathbb{Z}_2 \oplus \mathbb{Z}_4$.

Case 2 If, instead, all order-4 elements of G square to give z then everything in the quotient G/H has order at most 2, and so G/H is abelian by Proposition 1.20. Let $x \in G$ be an element of order 4. Since G/H is abelian, every conjugate of x in G must lie in the coset Hx .

³⁴ M Wild, *The groups of order sixteen made easy*, The American Mathematical Monthly 112 (2005) 20–31.

The conjugacy class $[x]$ of x thus has at most $|Hx| = 2$ elements, and by Corollary 6.27 the centraliser $Z_G(x)$ has order at least 8. We can therefore find an element $y \in Z_G(x) \setminus \langle x \rangle$, which commutes with x .

We want to use the elements x and y to generate a subgroup isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_4$. If y has order 2 then $\langle x, y \rangle$ is just such a group. If, however, $|y| = 4$ then by the above hypothesis $y^2 = z$ and we have $(xy)^2 = x^2y^2 = z^2 = e$, so $|xy|$ is at most 2. But $|xy|$ can't equal 1, because then we would have $xy = e$ and $y = x^{-1} \in \langle x \rangle$, which contradicts our choice of y , so $|xy| = 2$. Furthermore, $xy \notin \langle x \rangle$ because then we would have $xy = x^2 = e$, and again $y = x^{-1} \in \langle x \rangle$. So the subgroup $\langle x, xy \rangle$ must be isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_4$ as claimed; again it has index 2 and is therefore normal in G . $\square$

Every group of order 16, then, can be constructed as an extension of $\mathbb{Z}_2$ by one of three groups of order 8: either $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, or $\mathbb{Z}_2 \oplus \mathbb{Z}_4$, or the cyclic group $\mathbb{Z}_8$. By the Cyclic Extension Theorem,³⁵ to obtain all the groups of order 16 apart from $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, we just have to construct a list of cyclic extension types of the form $(\mathbb{Z}_8, 2, \tau, v)$ or $(\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \tau, v)$. The Cyclic Extension Theorem ensures that all of these are realised by valid groups, and then all we need to do is discard any duplicates on the resulting list.

The automorphisms of $\mathbb{Z}_8 = \langle x : x^8 = 1 \rangle$ are completely determined by their action on the generator x ; we list these in Table 7.8.

Suppose that $H = \langle x : x^8 = 1 \rangle = \mathbb{Z}_8$ is normal in G , and consider a non-identity element $g \in G \setminus H$ of minimal order. By Proposition 2.34 this must have order 2, 4, 8 or 16.

Case 1 ($|g| = 2$) In this case $v = g^2 = e$. All four automorphisms in Table 7.8 fix e and satisfy the requirement that τ^2 is conjugation by e . We therefore get four cyclic extension types $(\mathbb{Z}_8, 2, \phi_i, e)$ for $1 \leq i \leq 4$, which are realised by the following groups:

$$\begin{aligned} (\mathbb{Z}_8, 2, \phi_1, e) &\cong \mathbb{Z}_8 \oplus \mathbb{Z}_2, & (\mathbb{Z}_8, 2, \phi_2, e) &\cong \mathbb{Z}_8 \rtimes_{\phi_2} \mathbb{Z}_2 \cong SD_{16}, \\ (\mathbb{Z}_8, 2, \phi_3, e) &\cong \mathbb{Z}_8 \rtimes_{\phi_3} \mathbb{Z}_2, & (\mathbb{Z}_8, 2, \phi_4, e) &\cong \mathbb{Z}_8 \rtimes_{\phi_4} \mathbb{Z}_2 \cong D_8. \end{aligned}$$

The second of these is the **semidihedral group** SD_{16} of order 16.³⁶

Case 2 ($|g| = 4$) Here $v = g^2$ must have order 2; the only possibility is that $v = x^4$. If $\tau = \phi_1$ then $(x^2g)(x^2g) = x^2\phi_1(x^2)g^2 = x^4g^2 = x^8 = e$, so $|x^2g| = 2$ which contradicts the hypothesis that g has minimal order in $G \setminus H$. If $\tau = \phi_2$ then xg has order 2 and if $\tau = \phi_3$ then x^2g again has order 2. So the only new cyclic extension type is $(\mathbb{Z}_8, 2, \phi_4, x^2)$, which is realised by the dicyclic group Dic_4 .

Case 3 ($|g| = 8$) In this case there are two possible choices for v , namely x^2 and x^6 . Both of these elements are fixed by ϕ_1 and ϕ_3 , but not ϕ_2 or ϕ_4 , and so we have four potentially new cyclic extension

³⁵ Theorem 7.101, page 288.

ϕ	$\phi(x)$
$\phi_1 = \text{id}$	x
ϕ_2	x^3
ϕ_3	x^5
ϕ_4	$x^7 = x^{-1}$

Table 7.8: Automorphisms of $\mathbb{Z}_8$

³⁶ Some books call this the **quasidihedral group** QD_{16} ; it has presentation

$$\langle x, y : x^8 = y^2 = 1, yxy^{-1} = x^3 \rangle.$$

This is one of a family of groups SD_{2^n} or QD_{2^n} of the form

$$\langle x, y : x^{2^{n-1}} = y^2 = 1, yxy^{-1} = x^{2^{n-2}-1} \rangle$$

for $n \geq 2$. The case $n = 4$ is the first nonabelian group in the sequence, and the first that we haven't met in other guises, since $SD_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$ and $SD_8 \cong \mathbb{Z}_4 \oplus \mathbb{Z}_2$.

types

$$(\mathbb{Z}_8, 2, \phi_1, x^2), (\mathbb{Z}_8, 2, \phi_1, x^6), (\mathbb{Z}_8, 2, \phi_3, x^2) \text{ and } (\mathbb{Z}_8, 2, \phi_3, x^6).$$

The automorphism ϕ_4 maps x^2 to x^6 and commutes with both ϕ_1 and ϕ_3 , so the first and second of the above types are equivalent, as are the third and fourth.

Considering $(\mathbb{Z}_8, 2, \phi_1, x^2)$ we note that $(x^3g)(x^3g) = x^3\phi_1(x^3)g^2 = x^8 = e$. Therefore $|x^3g| = 2$ and so this extension type doesn't yield a group we haven't already seen. We can discard $(\mathbb{Z}_8, 2, \phi_3, x^2)$ as well, because $(xg)(xg) = x\phi_3(x)g^2 = x^8 = e$ and so $|xg| = 2$.

Case 4 ($|g| = 16$) If g has order 16 then the cyclic subgroup $\langle g \rangle$ must be the whole of G , so $G \cong \mathbb{Z}_{16}$. This group realises, amongst others, the cyclic extension type $(\mathbb{Z}_8, 2, \phi_1, x)$.

We have now found six groups of order 16 that have a normal subgroup isomorphic to $\mathbb{Z}_8$. Next we have to study the cyclic extension types of the form $(\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \tau, v)$. The automorphism group of $\mathbb{Z}_2 \oplus \mathbb{Z}_4 = \langle x, y : x^4 = y^2 = 1, xy = yx \rangle$ is isomorphic to the dihedral group D_8 . Table 7.9 lists the eight possible automorphisms and their actions on the generators x and y . With Proposition 7.104 in mind, we note that ψ_5 and ψ_7 are conjugate in $\text{Aut}(H)$ and so are ψ_6 and ψ_8 . We can therefore discard ψ_7 and ψ_8 , as they will give rise to groups isomorphic to those obtained from, respectively, ψ_5 and ψ_6 .

Also, since H is abelian, conjugation by any element of H is the identity map, so we can also rule out ψ_2 and ψ_4 , neither of which satisfy the requirement $\tau^2 = \text{id}$.

We assume now that G has no elements of order 8 or higher, otherwise it would have a normal subgroup isomorphic to $\mathbb{Z}_8$ and have been covered already in one of the previous four cases. Let $H = \langle x, y : x^4 = y^2 = 1, xy = yx \rangle \cong \mathbb{Z}_4 \oplus \mathbb{Z}_2$ and suppose that $g \in G \setminus H$ is a non-identity element of minimal order.

Case 5 ($|g| = 2$) In this case $v = g^2 = e$, which is fixed by any of the remaining four automorphisms under consideration. We thus obtain four cyclic extension types

$$\begin{aligned} (\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \phi_1, e) &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_4, & (\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \phi_3, e) &\cong D_4 \times \mathbb{Z}_2, \\ (\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \phi_5, e) &\cong V_4 \rtimes \mathbb{Z}_4, & (\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \phi_6, e) &\cong Q_8 \rtimes \mathbb{Z}_2. \end{aligned}$$

Case 6 ($|g| = 4$) Here v can be either x^2 , y or x^2y . From now on we assume that all elements of $G \setminus H$ have order at least 4.

Case 6a ($v = x^2$) If $v = x^2$ then only ψ_3 and ψ_5 determine groups that we haven't already seen. For ψ_1 the element xg has order 2, since $(xg)(xg) = x^2g^2 = x^4 = e$, and for ψ_6 the element xyg has order 2, since $(xyg)(xyg) = xyx^5yg^2 = x^2y^2g^2 = x^4y^2 = e$. Hence this case

ψ	$\psi(x)$	$\psi(y)$
$\psi_1 = \text{id}$	x	y
ψ_2	x^3y	x^2y
ψ_3	x^3	y
ψ_4	xy	x^2y
ψ_5	xy	y
ψ_6	x^3	x^2y
ψ_7	x^3y	y
ψ_8	x	x^2y

Table 7.9: Automorphisms of $\mathbb{Z}_2 \oplus \mathbb{Z}_4$

gives two new groups:

$$(\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \psi_3, x^2) \cong Q_8 \times \mathbb{Z}_2, \quad (\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \psi_5, x^2) \cong \mathbb{Z}_4 \rtimes \mathbb{Z}_4.$$

Case 6b ($v = y$) Here the condition $\tau(v) = v$ immediately rules out ψ_2, ψ_4, ψ_6 and ψ_8 . The identity automorphism $\psi_1 = \text{id}$ yields a new group

$$(\mathbb{Z}_2 \oplus \mathbb{Z}_4, 2, \psi_1, y) \cong \mathbb{Z}_4 \oplus \mathbb{Z}_4.$$

For ψ_3 we observe that $gx^2g^{-1} = \psi_3(x^2) = x^2$, so x^2 and g commute, and hence $\langle g, x^2 \rangle \cong \mathbb{Z}_2 \oplus \mathbb{Z}_4$. Furthermore, $(xg)(xg) = x\psi_3(x)g^2 = x^4y = y$, so if we replace g with $xg \notin \langle g, x^2 \rangle$, we have Case 6a again.

For ψ_5 we have $(xg)(xg) = x\psi_5(x)g^2 = x^2y^2 = x^2$, and replacing g with xg also returns us to Case 6a.

And for ψ_7 we have $(xg)(xg) = x\psi_7(x)g^2 = x^4yg^2 = x^4y^2 = e$, so $|xg| = 2$, which was already dealt with in Case 5.

Case 6c ($v = x^2y$) Finally, the condition $\tau(v) = v$ enables us to discard ψ_2, ψ_4, ψ_6 and ψ_8 straight away. There exists an automorphism of H (ψ_2 , for example) mapping $v = x^2y$ to y , and the allowed automorphisms ψ_1, ψ_3, ψ_5 and ψ_7 are closed under conjugation, so by Proposition 7.103 we get the same groups as in Case 6b.

We now have fourteen groups of order 16, namely the direct sum $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, together with six extensions of $\mathbb{Z}_2$ by $\mathbb{Z}_8$, and seven extensions of $\mathbb{Z}_2$ by $\mathbb{Z}_2 \oplus \mathbb{Z}_4$. It is not immediately obvious that some of these aren't isomorphic to each other, although this can be checked by comparing various fundamental properties such as commutativity and the number of elements of a given order. We will omit this, encouraging the interested reader to check it themselves. The fourteen groups of order 16 are listed in Table 7.10. (Several of these can also be constructed as extensions of $\mathbb{Z}_2$ by $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$, by D_4 , or by Q_8 .)

(i)	The cyclic group $\mathbb{Z}_{16}$,
(ii)	the abelian groups $\mathbb{Z}_2 \oplus \mathbb{Z}_8$,
(iii)	$\mathbb{Z}_4 \oplus \mathbb{Z}_4$,
(iv)	$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_4$ and
(v)	$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$,
(vi)	the dihedral group $D_8 \cong \mathbb{Z}_8 \rtimes_{\phi_4} \mathbb{Z}_2$,
(vii)	the dicyclic group Dic_4 ,
(viii)	the semidihedral group $SD_{16} \cong \mathbb{Z}_8 \rtimes_{\phi_2} \mathbb{Z}_2$,
(ix)	the direct products $D_4 \times \mathbb{Z}_2$ and
(x)	$Q_8 \times \mathbb{Z}_2$, and
(xi)	the semidirect products $\mathbb{Z}_8 \rtimes_{\phi_3} \mathbb{Z}_2$,
(xii)	$V_4 \rtimes \mathbb{Z}_4$,
(xiii)	$Q_8 \rtimes \mathbb{Z}_2$ and
(xiv)	$\mathbb{Z}_4 \rtimes \mathbb{Z}_4$.

Table 7.10: Groups of order 16

Groups of order 18

Let G be a group of order 18. By Sylow's Theorems G has exactly one Sylow 3-subgroup H of order 9, which is therefore normal in G by Corollary 7.12. Furthermore, there can be either 1, 3 or 9 Sylow 2-subgroups of order 2, each isomorphic to $\mathbb{Z}_2$.

Let $K \langle z : z^2 = 1 \rangle \cong \mathbb{Z}_2$ be a Sylow 2-subgroup of order 2. Then $G = HK$, an internal semidirect product of H with K . By Proposition 7.20, H is isomorphic to either $\mathbb{Z}_9$ or $\mathbb{Z}_3 \oplus \mathbb{Z}_3$.

Case 1 Let $H = \langle x : x^9 = 1 \rangle \cong \mathbb{Z}_9$. By Proposition 7.107, the semidirect product is defined by the action of K on H , and hence $zxz^{-1} = x^i$ for some i where $1 \leq i < 9$ and $i^2 \equiv 1 \pmod{9}$. There are two possible cases: $i = 1$ or $i = 8$.

Case 1a ($i = 1$) This is the trivial action, so $G \cong \mathbb{Z}_9 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_{18}$.

Case 1b ($i = 8$) This gives $zxz^{-1} = x^8 = x^{-1}$, and hence

$$G = \langle x, z : x^9 = 1, z^2 = 1, zxz^{-1} = x^{-1} \rangle$$

which is isomorphic to the dihedral group D_9 by Proposition 5.28.

Case 2 Let $H = \langle x, y : x^3 = 1, y^3 = 1, xy = yx \rangle = \langle x : x^3 = 1 \rangle \oplus \langle y : y^3 = 1 \rangle$. To determine the action of K on H we need to define how conjugation by z affects the generators x and y .

Suppose that $zxz^{-1} = t$ and that $t \notin \langle x \rangle$. Then $x = z^2xz^{-2} = ztz^{-1}$, so $zx = tz$ and $xz = zt$. Furthermore, $z(xt)z^{-1} = (zxz^{-1})(ztz^{-1}) = tx = xt$. If $t \notin \langle x \rangle$ then $H = \langle xt, x \rangle = \langle xt \rangle \oplus \langle x \rangle$. So, we can replace x with xt and y with x to get an alternative presentation for H where $zxz^{-1} = x$. We can therefore assume $zxz^{-1} \in \langle x \rangle$, and hence that $zxz^{-1} = x$ or x^{-1} .

Now $zyz^{-1} = x^i y^j$ for some i and j with $0 \leq i < 3$ and $0 < j < 3$. Hence $zx^i y^j z^{-1} = y$.

If $j = 1$ then $zyz^{-1} = x^i y$ and

$$y = z^2 y z^{-2} = zx^i y z^{-1} = (zx^i z^{-1})(zyz^{-1}) = (x^i)(x^i y) = x^{2i} y,$$

so $2i \equiv 0 \pmod{3}$, and hence $i = 0$. Thus if $j = 1$ then $zyz^{-1} = y$.

If $j = 2$ then $zyz^{-1} = x^i y^2$ and

$$y = z^2 y z^{-2} = zx^i y^2 z^{-1} = (zx^i z^{-1})(zy^2 z^{-1}) = (x^i)(x^{2i} y^4) = x^{3i} y.$$

Thus $3i \equiv 0 \pmod{3}$ and so $i = 0, 1, 2$. If $i = 0$ then $zyz^{-1} = y^2 = y^{-1}$; if $i = 1$ then $z(xy)z^{-1} = (xy)^{-1}$; and if $i = 2$ then $z(x^2 y)z^{-1} = (x^2 y)^{-1}$. In the latter two cases we can carefully redefine y to get $zyz^{-1} = y^{-1}$.

Otherwise, suppose that $zxz^{-1} = x^{-1}$. Then

$$y = zx^i y^j z^{-1} = x^{-i} x^{ij} y^{j^2} = x^{(i-1)j} y^{j^2}.$$

In this case, if $j = 1$ then by suitably redefining y we get $zyz^{-1} = y^{-1}$, and if $j = 2$ then $i = 0$ and $zyz^{-1} = y^2 = y^{-1}$.

We can therefore assume that $zxz^{-1} = x$ or x^{-1} , and $zyz^{-1} = y$ or y^{-1} .

Case 2a ($zxz^{-1} = x$ and $zyz^{-1} = y$) These are both trivial actions, so $xz = zx$ and $yz = zy$ and hence $G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_3 \oplus \mathbb{Z}_6$.

Case 2b ($zxz^{-1} = x$ and $zyz^{-1} = y^{-1}$) In this case, we have

$$G = \langle x : x^3 = 1 \rangle \oplus \langle y, z : y^3 = z^2 = 1, zyz^{-1} = y^{-1} \rangle \cong \mathbb{Z}_3 \times D_3.$$

Case 2c ($zxz^{-1} = x^{-1}$ and $zyz^{-1} = y^{-1}$) This gives the group with presentation

$$G = \langle x, y, z : x^3 = y^3 = z^2 = 1, zxz^{-1} = x^{-1}, zyz^{-1} = y^{-1} \rangle.$$

This is a nontrivial semidirect product $(\mathbb{Z}_3 \oplus \mathbb{Z}_3) \rtimes \mathbb{Z}_2$.

(i)	the cyclic group $\mathbb{Z}_{18} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_9$,
(ii)	the abelian group $\mathbb{Z}_3 \oplus \mathbb{Z}_6 \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_2$,
(iii)	the dihedral group D_9 ,
(iv)	the direct product $\mathbb{Z}_3 \times D_3$, and
(v)	the semidirect product $(\mathbb{Z}_3 \oplus \mathbb{Z}_3) \rtimes \mathbb{Z}_2$.

Table 7.11: Groups of order 18

All that remains is to show that these five groups are nonisomorphic. The first and third are abelian and distinct by Theorem 5.48. The dihedral group D_9 has an element x of order 9, whereas neither $\mathbb{Z}_3 \times D_3$ nor $(\mathbb{Z}_3 \oplus \mathbb{Z}_3) \rtimes \mathbb{Z}_2$ do. Finally, $\mathbb{Z}_3 \times D_3$ has an element xz of order 6, but $(\mathbb{Z}_3 \oplus \mathbb{Z}_3) \rtimes \mathbb{Z}_2$ has no elements of order greater than 3.

Therefore, up to isomorphism, there are five groups of order 18, listed in Table 7.11

Groups of order 20

Let G be a group of order 20. By Sylow's Theorems there is a single Sylow 5-subgroup $H = \langle x : x^5 = 1 \rangle$ of order 5, which is normal in G , and which must be isomorphic to $\mathbb{Z}_5$. There must be 5 Sylow 2-subgroups of order 4, which must be isomorphic to either $\mathbb{Z}_4$ or $\mathbb{Z}_2 \oplus \mathbb{Z}_2$. Let K be one of these Sylow 2-subgroups. Then $H \cap K = \{e\}$ and $HK = G$, so G is an internal semidirect product of H with K .

Case 1 Let $K = \langle y : y^4 = 1 \rangle \cong \mathbb{Z}_4$. This acts on H by conjugation, so we need to investigate the possible values of $yx y^{-1} = x^i$, where $1 \leq i \leq 4$. Also, $x = y^4 x y^{-4} = x^{i^4}$, so $i \equiv 1 \pmod{5}$. All of $i = 1, 2, 3, 4$ satisfy these conditions, but we can discard $i = 3$, because if $yx y^{-1} = x^2$ then $y^3 x y^{-3} = x^{2^3} = x^8 = x^3$, and since $\langle y \rangle = \langle y^3 \rangle$, the group obtained by setting $i = 2$ is isomorphic to the one we get by setting $i = 3$.

Case 1a ($i = 1$) Here $yx y^{-1} = x$, so

$$G = \langle x, y : x^5 = y^4 = 1, yx = xy \rangle \cong \mathbb{Z}_5 \oplus \mathbb{Z}_4 \cong \mathbb{Z}_{20}.$$

Case 1b ($i = 2$) In this case, $yx y^{-1} = x^2$, giving the presentation

$$G = \langle x, y : x^5 = y^4 = 1, yx = x^2 y \rangle.$$

This is isomorphic to the affine general linear group $AGL_1(5)$ of affine linear transformations in the finite field $\mathbb{F}_5$.

Case 1c ($i = 4$) This time, $yx y^{-1} = x^4 = x^{-1}$, which gives us the presentation

$$G = \langle x, y : x^5 = y^4 = 1, yx y^{-1} = x^{-1} \rangle.$$

This group is isomorphic to the dicyclic group Dic_5 . It is not isomorphic to the group in Case 1b, because that group has trivial centre, whereas here $y^2 x y^{-2} = x$, so $y^2 x = x y^2$, and thus y^2 is contained in the centre.

Case 2 Let $K = \langle y, z : y^2 = z^2 = 1, yz = zy \rangle \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. The K -action on H is determined by the values of $yx y^{-1}$ and $zx z^{-1}$. As before, suppose that $yx y^{-1} = x^i$ for some i with $1 \leq i \leq 4$. Then $x = y^2 x y^{-2} = x^{i^2}$, and so $i \equiv 1 \pmod{5}$. The only possibilities are

$i = 1, 4$, so either $xyx^{-1} = x$ or $xyx^{-1} = x^{-1}$. Similarly, $z x z^{-1} = x$ or $z x z^{-1} = x^{-1}$. For reasons analogous to those in case 1b of the order 12 classification, we may assume that $xyx^{-1} = x$, and we therefore have two cases to consider.

Case 2a ($z x z^{-1} = x$) This gives the presentation

$$G = \langle x, y, z : x^5 = y^2 = z^2 = 1, xy = yx, xz = zx, yz = zy \rangle$$

which is isomorphic to the abelian group $\mathbb{Z}_5 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_{10}$.

Case 2b ($z x z^{-1} = x^{-1}$) This gives the presentation

$$G = \langle x, y, z : x^5 = y^2 = z^2 = 1, yx = xy, zx = x^{-1}z, yz = zy \rangle \\ \cong \langle x, z : x^5 = z^2 = 1, zx = x^{-1}z \rangle \times \langle y : y^2 = 1 \rangle,$$

which is isomorphic to the direct product $D_5 \times \mathbb{Z}_2$, which in turn can be shown to be isomorphic to the dihedral group D_{10} .

There are thus five isomorphism classes of groups with 20 elements, listed in Table 7.12.

(i)	The cyclic group $\mathbb{Z}_{20} \cong \mathbb{Z}_4 \oplus \mathbb{Z}_5$,
(ii)	the abelian group $\mathbb{Z}_2 \oplus \mathbb{Z}_{10} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_5$,
(iii)	the dihedral group D_{10} ,
(iv)	the dicyclic group Dic_5 , and
(v)	the affine general linear group $AGL_1(5)$.

Table 7.12: Groups of order 20

Groups of order 21

Let G be a group of order 21. Then by Sylow's Theorems there is a unique Sylow 7-subgroup $H = \langle x : x^7 = 1 \rangle \cong \mathbb{Z}_7$, which is normal in G . Suppose that $K = \langle y : y^3 = 1 \rangle \cong \mathbb{Z}_3$ is a Sylow 3-subgroup. Then $H \cap K$ is trivial and $HK = G$, so G is an internal semidirect product.

Considering $yx y^{-1} = x^i$, with $1 \leq i \leq 6$, then $x = y^3 x y^{-3} = x^{i^3}$, so $i^3 \equiv 1 \pmod{7}$. The possible values for i are thus $i = 1, 2$ or 4 .

If $yx y^{-1} = x^2$ then $y^2 x y^{-2} = x^4$, and since $\langle y \rangle = \langle y^2 \rangle$ we can replace y with y^2 and see that the cases $i = 2$ and $i = 4$ give isomorphic groups.

Case 1 ($i = 1$) This gives the presentation

$$G = \langle x, y : x^7 = y^3 = 1, xy = yx \rangle \cong \mathbb{Z}_7 \oplus \mathbb{Z}_3 \cong \mathbb{Z}_{21}.$$

Case 2 ($i = 2$) Here,

$$G = \langle x, y : x^7 = y^3 = 1, yx = x^2 y \rangle,$$

which is a nontrivial semidirect product $\mathbb{Z}_7 \rtimes \mathbb{Z}_3$.

There are thus two nonisomorphic groups of order 21, listed in Table 7.13.

(i)	The cyclic group $\mathbb{Z}_{21} \cong \mathbb{Z}_3 \oplus \mathbb{Z}_7$ and
(ii)	the nontrivial semidirect product $\mathbb{Z}_7 \rtimes \mathbb{Z}_3$.

Table 7.13: Groups of order 21

Groups of order 24

Let G be a group of order 24. By Sylow's Theorems, there exist either 1 or 3 Sylow 2-subgroups of order 8, and either 1 or 4 Sylow 3-subgroups of order 3.

Suppose that $n_3 = 4$. By Sylow's Second Theorem, these four Sylow 3-subgroups are conjugate to each other. That is, G acts on $\text{Syl}_3(G)$ by conjugation and thereby determines a homomorphism $f: G \rightarrow S_4$, where each $g \in G$ maps to the permutation of $\text{Syl}_3(G)$ defined by the conjugation action.

By Sylow's Second Theorem there must exist elements of G that map each of the four Sylow 3-subgroups to any of the others. So $\text{im}(f)$ must have at least 4 elements and hence $|\ker(f)| \leq 6$. By Lagrange's Theorem, $|\ker(f)|$ can be 6, 3, 2 or 1.

If $\ker(f)$ contains one Sylow 3-subgroup then it must contain all of them: $\ker(f)$ is normal in G and hence closed under conjugation by all elements of G , and by Sylow's Second Theorem the Sylow 3-subgroups are conjugate, so either all of them must lie in $\ker(f)$ or none of them do. The first of these is ruled out by the order of $\ker(f)$, which has at most six elements and is thus not big enough to contain four distinct 3-element subgroups. So $|\ker(f)|$ can't be 3 or 6. If $|\ker(f)| = 1$ then f is an isomorphism and hence $G \cong S_4$. This has four Sylow 3-subgroups and three Sylow 2-subgroups.

If $|\ker(f)| = 2$ then by the First Isomorphism Theorem, $\text{im}(f) \cong G/\ker(f)$ is a subgroup of S_4 of order 12; it can only be the alternating group A_4 , which has a single Sylow 2-subgroup of order 4. Therefore G itself can only have one Sylow 2-subgroup (of order 8) and hence $n_2 = 1$.

Case 1 ($n_2 = 1, n_3 = 1$) In this case, both the Sylow 2-subgroup H and the Sylow 3-subgroup K are normal in G . Their intersection is trivial and $G = HK$, so $G \cong H \times K$. There are five possibilities for H and one for K , so this case yields five groups for our list:

$$\begin{aligned} \mathbb{Z}_8 \oplus \mathbb{Z}_3 &\cong \mathbb{Z}_{24}, & D_4 \times \mathbb{Z}_3, \\ \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_3 &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_{12}, & Q_8 \times \mathbb{Z}_3, \\ \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3 &\cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6. \end{aligned}$$

Case 2 ($n_2 = 1, n_3 = 4$) Here we have a normal Sylow 2-subgroup H of order 8 and four conjugate Sylow 3-subgroups. Let K be one of these subgroups. Then $H \cap K = \{e\}$ and we have a semidirect product $G = HK$. We have five choices for H and must consider the possible $\mathbb{Z}_3$ -actions. The trivial action yields the five direct products already listed in Case 1, so at this point we're only interested in nontrivial $\mathbb{Z}_3$ -actions.

Case 2a ($H \cong \mathbb{Z}_8$) Let $H = \langle x : x^8 = 1 \rangle$ and $K = \langle w : w^3 = 1 \rangle$. By Proposition 7.107 we have $wxw^{-1} = x^i$ with $0 < i < 8$ and $i^3 \equiv 1 \pmod{8}$. The only value of i satisfying these criteria is $i = 1$, which yields the trivial action.

Case 2b ($H \cong \mathbb{Z}_2 \oplus \mathbb{Z}_4$) We want a nontrivial homomorphism $\phi: K \rightarrow \text{Aut}(H)$ mapping the generator of K to a nontrivial element of $\text{Aut}(H)$. The generator of K has order 3, so its image in $\text{Aut}(H)$ must also have order 3; however by Proposition 2.34 the order of any element of $\text{Aut}(H)$ must divide $|\text{Aut}(H)| = 8$, hence no such homomorphism ϕ can exist. This case also only gives the trivial action and one of the direct products we've already seen.

Case 2c ($H \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$) Again, we want a nontrivial homomorphism $\phi: K \rightarrow \text{Aut}(H)$. This time $|\text{Aut}(H)| = 168$ (it happens to be isomorphic to the groups $GL_3(2)$ and $PSL_2(7)$) which is divisible by 3. By Cauchy's Theorem, it therefore has at least one subgroup isomorphic to $\mathbb{Z}_3$, whose generator has order 3. In fact, all such elements are conjugate to each other, so we have only a single nontrivial action at our disposal: the map that cyclically permutes the factors of $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$. This action yields a new group, a nontrivial semidirect product $G = (\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2) \rtimes \mathbb{Z}_3$, which is isomorphic to the direct product $\mathbb{Z}_2 \times A_4$.

Case 2d ($H \cong D_4$) There is no nontrivial $\mathbb{Z}_3$ -action on D_4 since $|\text{Aut}(D_4)| = 8$, so we don't get any new groups in this case.

Case 2e ($H \cong Q_8$) In this case $\text{Aut}(Q_8) \cong S_4$ and thus $|\text{Aut}(Q_8)| = 24$, which is divisible by 3, so we can define a nontrivial $\mathbb{Z}_3$ -action on Q_8 . All of the 3-element subgroups of S_4 are conjugate to each other, so we have only one nontrivial action, which yields a new group for our list: the nontrivial semidirect product $G = Q_8 \rtimes \mathbb{Z}_3$. This is isomorphic to the group $SL_2(3)$ of 2×2 matrices over $\mathbb{F}_3$ with determinant 1.

Case 3 ($n_2 = 3, n_3 = 1$) If G has a single Sylow 3-subgroup $H \cong \mathbb{Z}_3$ and K is one of the three Sylow 2-subgroups of order 8, then again we have a semidirect product $G = HK$ whose structure is determined by the K -action on H . As before, there are five choices for K , and by this stage we are only interested in nontrivial K -actions: nontrivial homomorphisms $\phi: K \rightarrow \text{Aut}(\mathbb{Z}_3) = \mathbb{Z}_2$.

Case 3a ($K = \mathbb{Z}_8$) There is only one nontrivial action of $\mathbb{Z}_8$ on $\mathbb{Z}_3$, given by the homomorphism $\phi: \mathbb{Z}_8 \rightarrow \mathbb{Z}_2 = \text{Aut}(\mathbb{Z}_3)$ with kernel $\{0, 2, 4, 6\} \cong \mathbb{Z}_4$. This yields a semidirect product $\mathbb{Z}_3 \rtimes \mathbb{Z}_8$.

Case 3b ($K = \mathbb{Z}_2 \oplus \mathbb{Z}_4$) Here we need to study actions of $\mathbb{Z}_2 \oplus \mathbb{Z}_4$ on $\mathbb{Z}_3$, which are determined by nonzero homomorphisms $\phi: \mathbb{Z}_2 \oplus \mathbb{Z}_4 \rightarrow \mathbb{Z}_2$. There are essentially two such homomorphisms, with kernels isomorphic to $\mathbb{Z}_4$ or the Klein group V_4 . The first of these yields a semidirect product $\mathbb{Z}_3 \rtimes (\mathbb{Z}_2 \oplus \mathbb{Z}_4) \cong \mathbb{Z}_4 \times D_3$, while the second gives $\mathbb{Z}_3 \rtimes (\mathbb{Z}_2 \oplus \mathbb{Z}_4) \cong \text{Dic}_3 \times \mathbb{Z}_2$.

(i)	The cyclic group $\mathbb{Z}_{24} \cong \mathbb{Z}_3 \oplus \mathbb{Z}_8$,
(ii)	the abelian groups $\mathbb{Z}_2 \oplus \mathbb{Z}_{12} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_3$ and
(iii)	$\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6 \cong$ $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6$,
(iv)	the symmetric group S_4 ,
(v)	the dihedral group D_{12} ,
(vi)	the direct products $\mathbb{Z}_3 \times D_4$,
(vii)	$\mathbb{Z}_3 \times Q_8$,
(viii)	$\mathbb{Z}_2 \times A_4$,
(ix)	$\mathbb{Z}_4 \times D_3$,
(x)	$\mathbb{Z}_2 \times \text{Dic}_3$
(xi)	and $\mathbb{Z}_2 \times D_6 \cong V_4 \times D_3$,
(xii)	and the semidirect products $Q_8 \rtimes \mathbb{Z}_3 \cong \text{SL}_2(3)$,
(xiii)	$\mathbb{Z}_3 \rtimes \mathbb{Z}_8$,
(xiv)	$\mathbb{Z}_3 \rtimes D_4$ (with kernel V_4)
(xv)	and $\mathbb{Z}_3 \rtimes Q_8$.

Table 7.14: Groups of order 24

Case 3c ($K = \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2$) In this case, we want a nonzero homomorphism $\phi: \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \rightarrow \mathbb{Z}_2$, of which there is only one, with kernel isomorphic to V_4 . This action gives a semidirect product $\mathbb{Z}_3 \rtimes (\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2)$, which is isomorphic to the direct products $\mathbb{Z}_2 \times D_6 \cong D_3 \times V_4$.

Case 3d ($K = D_4$) There are two nontrivial D_4 -actions on $\mathbb{Z}_3$, determined by homomorphisms $\phi: D_4 \rightarrow \mathbb{Z}_2$. The one with kernel $R_4 = \langle r \rangle \cong \mathbb{Z}_4$ yields a semidirect product $\mathbb{Z}_3 \rtimes D_4$ which is isomorphic to the dihedral group D_{12} . The one with kernel isomorphic to V_4 yields another group $\mathbb{Z}_3 \rtimes D_4$.

Case 3e ($K = Q_8$) There is just one nontrivial Q_8 -action on $\mathbb{Z}_3$, determined by the homomorphism $\phi: Q_8 \rightarrow \mathbb{Z}_2$ with kernel $\langle I \rangle, \langle J \rangle$ or $\langle K \rangle$, all of which are isomorphic to the Klein group V_4 . This yields a semidirect product $\mathbb{Z}_3 \rtimes Q_8$.

Case 4 ($n_2 = 3, n_3 = 4$) By the earlier discussion $G = S_4$.

There are thus fifteen groups of order 24, listed in Table 7.14.

Groups of order 27

Let G be a group of order $27 = 3^3$. It is a p -group, so by Proposition 7.19 it has nontrivial centre. By Lagrange's Theorem, $|Z(G)|$ can be either 3, 9 or 27. If the latter, then G is equal to its centre and hence abelian, so the classification theorem for finitely generated abelian groups yields three possible cases: $\mathbb{Z}_{27}$, $\mathbb{Z}_3 \oplus \mathbb{Z}_9$ and $\mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3$.

Now suppose that G is nonabelian and $|G| = 27$. For convenience let $Z = Z(G)$. Then $|Z| = 3$ or 9, which means that $|G/Z| = 9$ or 3, respectively. The latter case is ruled out by Proposition 7.108: if $|G/Z| = 3$ then $G/Z \cong \mathbb{Z}_3$, which can only happen if G is abelian. So $|G/Z| = 9$ and $|Z| = 3$, which means that $Z \cong \mathbb{Z}_3$, and $G/Z \cong \mathbb{Z}_9$ or $\mathbb{Z}_3 \oplus \mathbb{Z}_3$ by Proposition 7.20. But $G/Z \not\cong \mathbb{Z}_9$ because otherwise G would be abelian by Proposition 7.108. So $G/Z \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3$.

Since G/Z is abelian, the commutator subgroup $[G, G] \leq Z$, and since $Z \cong \mathbb{Z}_3$, a cyclic group of prime order, Lagrange's Theorem implies that $[G, G]$ is either trivial or equal to Z . But G is nonabelian, so $[G, G]$ can't be trivial, and therefore $[G, G] = Z \cong \mathbb{Z}_3$.

There are nontrivial elements $x, y \in G$ such that xZ and yZ generate G/Z , with neither xZ nor yZ being equal to the centre Z itself. These elements x and y don't commute: If they did, then the centraliser $C_G(x)$ would contain x, y and all of Z , which would mean that $C_G(x) = G$. This in turn would imply that everything commutes with x , so $x \in Z$ and $xZ = Z$, which contradicts our choice of x .

So there exists some nontrivial $z = [x, y] \in [G, G] = Z$; this element z has order 3 and thus generates $Z = \langle z \rangle$. The elements x and y must have order either 3 or 9: they're nontrivial so can't have order 1, and if either had order 27 then it would generate the entire group G , which would then be cyclic and abelian.

We therefore have three cases to consider:

Case 1 ($|x| = |y| = 3$) This gives the presentation

$$G = \langle x, y, z : x^3 = y^3 = z^3 = 1, xz = zx, yz = zy, [x, y] = z \rangle.$$

This is isomorphic to the **Heisenberg group** $Heis(3)$ or $U(3, 3)$ of **unitriangular** matrices over $\mathbb{F}_3$:

$$Heis(3) = U(3, 3) = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} : a, b, c \in \mathbb{F}_3 \right\}$$

Case 2 ($|x| = 9, |y| = 3$) In this case, observe that $x^3 \in Z$; since $x^3 \neq 1$ then it must generate Z and hence $x^3 = z$ or z^2 . This gives the presentation

$$G = \langle x, y : x^9 = y^3 = 1, x^3 = [x, y] \rangle.$$

This is isomorphic to the semidirect product $\mathbb{Z}_9 \rtimes \mathbb{Z}_3$.

Case 3 ($|x| = |y| = 9$) This time we get the presentations

$$\begin{aligned} G &= \langle x, y : x^9 = y^9 = 1, x^3 = y^3 = [x, y] \rangle \\ &\cong \langle x, y : x^9 = y^9 = 1, x^3 = [x, y], y^3 = [x, y]^2 \rangle, \end{aligned}$$

both of which are isomorphic to the group obtained in Case 2.

We therefore have five groups of order 27, listed in Table 7.15.

(i)	The cyclic group $\mathbb{Z}_{27}$,
(ii)	the abelian groups $\mathbb{Z}_3 \oplus \mathbb{Z}_9$ and
(iii)	$\mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3$,
(iv)	the Heisenberg group $Heis(3) = U(3, 3)$ of 3×3 unitriangular matrices over $\mathbb{F}_3$ and
(v)	the semidirect product $\mathbb{Z}_9 \rtimes \mathbb{Z}_3$.

Table 7.15: Groups of order 27

Groups of order 28

Let G be a group of order $28 = 2^2 \times 7$. Then G has a normal Sylow 7-subgroup $H = \langle x : x^7 = 1 \rangle \cong \mathbb{Z}_7$ and either 1 or 7 Sylow 2-subgroups of order 4. Let K be one of these latter subgroups, which can be isomorphic to either $\mathbb{Z}_4$ or $\mathbb{Z}_2 \oplus \mathbb{Z}_2$. The group G decomposes as an internal direct product HK , with K acting on H via a homomorphism $\phi: K \rightarrow \text{Aut}(H) \cong \mathbb{Z}_6$.

Case 1 ($K \cong \mathbb{Z}_4$) There are two possible actions of $K = \langle y : y^4 = 1 \rangle \cong \mathbb{Z}_4$ on H : the identity map given by $yx y^{-1} = x$, and the inversion map given by $yx y^{-1} = x^{-1}$. We therefore get two groups:

$$\begin{aligned} \langle x, y : x^7 = y^4 = 1, yx = xy \rangle &\cong \mathbb{Z}_4 \oplus \mathbb{Z}_7 \cong \mathbb{Z}_{28} \\ \text{and } \langle x, y : x^7 = y^4 = 1, yx = x^{-1}y \rangle &\cong \text{Dic}_7. \end{aligned}$$

Case 2 ($K \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$) Let $K = \langle y, z : y^2 = z^2 = 1, yz = zy \rangle$. There are three nontrivial homomorphisms $\phi: \mathbb{Z}_2 \oplus \mathbb{Z}_2 \rightarrow \text{Aut}(H) \cong \mathbb{Z}_6$, all with images either the identity map or the inversion map $x \mapsto x^{-1}$.

If both generators y and z map to the identity then we get the group

$$\langle x, y, z : x^7 = y^2 = z^2 = 1, xy = yx, xz = zx, yz = zy \rangle \\ \cong \mathbb{Z}_7 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_{14}.$$

If y maps to the inversion automorphism and z maps to the identity (or the other way round) then we get

$$\langle x, y, z : x^7 = y^2 = z^2 = 1, xy = yx^{-1}, xz = zx, yz = zy \rangle \\ \cong D_7 \times \mathbb{Z}_2 \cong D_{14}.$$

(i)	The cyclic group $\mathbb{Z}_{28} \cong \mathbb{Z}_4 \oplus \mathbb{Z}_7$,
(ii)	the abelian group $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_7 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_{14}$,
(iii)	the dicyclic group Dic_7 , and
(iv)	the dihedral group D_{14} .

If both y and z map to the inversion automorphism then we can replace either y or z with yz , which then maps to the identity, and this reduces to the previous case.

Table 7.16: Groups of order 28

Hence there are four groups of order 28, listed in Table 7.16.

Groups of order 30

Suppose that $|G| = 30 = 2 \times 3 \times 5$. Then by Sylow's Theorems and Lagrange's Theorem $n_3 = 1$ or 10 and $n_5 = 1$ or 6. At least one of n_3 and n_5 must be equal to 1: there aren't enough elements in G for there to be 10 distinct subgroups of order 3 and 6 distinct subgroups of order 5 with trivial intersections.

We now show that G contains a subgroup isomorphic to $\mathbb{Z}_{15}$. Since at least one of n_3 and n_5 equals 1, we can find subgroups $H \cong \mathbb{Z}_3$ and $K \cong \mathbb{Z}_5$, at least one of which is normal. These subgroups have trivial intersection and hence HK is a subgroup of order 15. This subgroup is isomorphic to $\mathbb{Z}_{15}$ by Example 7.14.

Now let $H = \langle x : x^{15} = 1 \rangle$ be this subgroup. It has index 2 in G and is hence normal by Proposition 3.12. By Sylow's Theorems and Lagrange's Theorem, G also contains at least one (and possibly 3 or 5) subgroups of order 2. Let $K = \langle y : y^2 = 1 \rangle$ be one of these subgroups; it has trivial intersection with H , and $HK = G$, so our group G decomposes as a semidirect product $\mathbb{Z}_{15} \rtimes \mathbb{Z}_2$. By Proposition 7.107 the required action is determined entirely by $yx y^{-1} = x^i$ for $0 < i < 15$ and $i^2 \equiv 1 \pmod{15}$. There are four cases to consider:

Case 1 ($i = 1$) This is the trivial action and yields the direct sum $\mathbb{Z}_{15} \oplus \mathbb{Z}_2 \cong \mathbb{Z}_{30}$.

Case 2 ($i = 4$) This case yields the presentation

$$G = \langle x, y : x^{15} = y^2 = 1, yxy^{-1} = x^4 \rangle.$$

This group is isomorphic to the direct product $D_5 \times \mathbb{Z}_3$.

Case 3 ($i = 11$) Here we obtain the presentation

$$G = \langle x, y : x^{15} = y^2 = 1, yxy^{-1} = x^{11} \rangle,$$

which is isomorphic to $D_3 \times \mathbb{Z}_5$.

Case 4 ($i = 14$) This final action gives the presentation

$$G = \langle x, y : x^{15} = y^2 = 1, yxy^{-1} = x^{-1} \rangle \cong D_{15}.$$

There are thus four different groups of order 30, listed in Table 7.17.

This completes our classification of groups of order less than 32. We stop here because things get considerably more complicated with groups of order 2^n , and in particular there are 51 groups of order 32. Table 7.18 lists the number of groups of order 2^n . To put this in perspective, up to isomorphism there are 49 910 529 484 groups with order at most 2000, and 49 487 365 422 of those (just over 99.15%) have order 1024.

(i)	The cyclic group $\mathbb{Z}_{30} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_5$,
(ii)	the dihedral group D_{15} and the direct products
(iii)	$D_3 \times \mathbb{Z}_5$ and
(iv)	$D_5 \times \mathbb{Z}_3$.

Table 7.17: Groups of order 30

	order	groups
2^0	= 1	1
2^1	= 2	1
2^2	= 4	2
2^3	= 8	5
2^4	= 16	14
2^5	= 32	51
2^6	= 64	267
2^7	= 128	2 328
2^8	= 256	56 092
2^9	= 512	10 494 213
2^{10}	= 1024	49 487 365 422

Table 7.18: Groups of order 2^n

Summary

Lagrange's Theorem³⁷ states that if G is a finite group then the order of any subgroup $H \leq G$ must divide the order of G . The converse is not true in general, the smallest counterexample being the alternating group A_4 , which has order 12 but no subgroup of order 6.³⁸ Cauchy's Theorem³⁹ provides a partial converse, however: if p is a prime factor of $|G|$ then G contains a cyclic subgroup of order p .

A finite group G is said to be a **p -group** if it has order p^k for some prime p and integer $k \geq 0$.⁴⁰ A subgroup of some (not necessarily finite) group G is a **p -subgroup** if it is itself a p -group.⁴¹

Suppose that G is a p -group acting on some finite set X . Then the number of elements in X is equal to the number of elements fixed by the G -action, modulo p . That is, $|X| \equiv |\text{Fix}_G(X)| \pmod{p}$.⁴²

We can use this fact to prove Sylow's First Theorem:⁴³ any finite group G contains at least one p -subgroup of each possible order for every prime factor p of $|G|$, and each of these subgroups is normal in the p -subgroup that contains it. That is, if G is a finite group of order $p^k m$, where p is prime, $k \geq 0$ and $p \nmid m$, then it has at least one subgroup of order p^i for all $0 \leq i \leq k$. Furthermore, each subgroup of order p^i is a normal subgroup of some subgroup of order p^{i+1} . A maximal p -subgroup is called a **Sylow p -subgroup**,⁴⁴ and Sylow's Second Theorem⁴⁵ says that for any prime factor p of $|G|$, the Sylow p -subgroups are conjugate to each other. Therefore, G has exactly one Sylow p -subgroup H if and only if H is normal.⁴⁶

³⁷ Theorem 2.30, page 54.

³⁸ Example 2.36, page 56.

³⁹ Theorem 2.37, page 57.

⁴⁰ Definition 7.1, page 234.

⁴¹ Definition 7.2, page 234.

⁴² Proposition 7.3, page 235.

⁴³ Theorem 7.4, page 236.

⁴⁴ Definition 7.10, page 239.

⁴⁵ Theorem 7.11, page 239.

⁴⁶ Corollary 7.12, page 239.

⁴⁷ Theorem 7.13, page 240.

⁴⁸ Example 7.14, page 241.

⁴⁹ Proposition 7.15, page 241.

⁵⁰ Corollary 7.16, page 242.

We denote by $\text{Syl}_p(G)$ the set of Sylow p -subgroups of G , and let $n_p = |\text{Syl}_p(G)|$ denote the number of Sylow p -subgroups. Sylow's Third Theorem⁴⁷ says that for any finite group G of order $p^k m$, we have $n_p \equiv 1 \pmod{p}$ and $n_p | m$. We can use this to help classify finite groups of a particular order: for example, if $|G| = 15$ we find that $n_3 = 1$ and $n_5 = 1$, so G must have a single normal subgroup of order 3 and another of order 5. From this we can deduce that $G \cong \mathbb{Z}_{15}$.⁴⁸ More generally, if $|G| = pq$ where p and q are distinct primes, $p < q$ and $q \not\equiv 1 \pmod{p}$ then $G \cong \mathbb{Z}_{pq}$.⁴⁹ Any such group must therefore be abelian.⁵⁰

We can also use Sylow's Second and Third Theorems to prove the nonexistence of finite simple groups of certain orders, by showing that a group of the given order must have a single Sylow p -subgroup for some p , and this subgroup must necessarily be normal. For example, there is no simple group of order 1246, since $1246 = 2 \cdot 7 \cdot 89$ and $n_7 = n_{89} = 1$.⁵¹

⁵¹ Example 7.18, page 242.

Another useful fact about p -groups is that they have nontrivial centre.⁵² We can use this to show that any group of order p^2 is isomorphic to either $\mathbb{Z}_{p^2}$ or $\mathbb{Z}_p \oplus \mathbb{Z}_p$,⁵³ and must in either case be abelian.⁵⁴

⁵² Proposition 7.19, page 243.

⁵³ Proposition 7.20, page 243.

⁵⁴ Corollary 7.21, page 244.

⁵⁵ Proposition 7.22, page 244.

More generally, if $|G| = p^2 q$ where p and q are both prime, $p < q$ and $q \not\equiv 1 \pmod{p}$, then G must be abelian.⁵⁵

We next discussed the concept of **subgroup series**: a finite nested sequence $\{e\} = H_0 < H_1 < \cdots < H_n = G$ of subgroups of a group G .⁵⁶ We sometimes make a distinction between an **ascending series** and a **descending series** depending on what order we number the subgroups in, and we will also sometimes relax the requirement that the smallest subgroup is trivial. A **normal series** or **invariant series** is one in which each subgroup H_i is normal in G , while a **subnormal series** satisfies the weaker condition that each subgroup H_i need only be normal in the next largest subgroup H_{i+1} .⁵⁷ Every normal series is subnormal, but not every subnormal series is normal.

⁵⁶ Definition 7.23, page 245.

⁵⁷ Definition 7.24, page 245.

The advantage of considering normal and subnormal series is that the quotients H_{i+1}/H_i are all well-defined. This approach provides a useful way of comparing series: two normal or subnormal series are **isomorphic** or **equivalent** if there is a bijective correspondence between their quotient groups.⁵⁸

⁵⁸ Definition 7.25, page 246.

Given any normal or subnormal series H of some group G , a **refinement** of H is a new series obtained by inserting additional subgroups.⁵⁹ **Schreier's Refinement Theorem**⁶⁰ says that given any two normal (or subnormal) series H and K for the same group G , we can find a refinement of H and a refinement of K that are isomorphic

⁵⁹ Definition 7.26, page 247.

⁶⁰ Theorem 7.27, page 248.

to each other. The proof of this is somewhat involved, and makes use of a rather technical result, **Zassenhaus' Lemma** or the **Butterfly Lemma**.^{61,62} Our proof of this required another technical lemma called **Dedekind's Modular Law**.⁶³

A **maximal subgroup** is a proper subgroup not contained in any larger proper subgroup, and a **maximal normal subgroup** is a proper normal subgroup not contained in any larger proper normal subgroup.⁶⁴ A quotient G/N is simple exactly when $N \triangleleft G$ is a maximal normal subgroup of G .⁶⁵ A **composition series** for a group G is a subnormal series that cannot be refined any further.⁶⁶ This is the same as requiring each group H_i to be a maximal normal subgroup of the next largest group H_{i+1} , or equivalently requiring each quotient (or **composition factor**) to be simple. A normal (rather than just subnormal) composition series is called a **principal series** or **chief series**. Not every group has a composition or principal series: in particular $\mathbb{Z}$ doesn't. The **Jordan–Hölder Theorem**⁶⁷ says that any two composition series for the same group are isomorphic. If G has a composition (or principal) series, then we can find one that includes any given normal subgroup $N \triangleleft G$.⁶⁸

We say that a group is **soluble** or **solvable** if its composition factors are all abelian;⁶⁹ this is equivalent to requiring the composition factors to be cyclic groups $\mathbb{Z}_p$ of prime order. The symmetric group S_n is not soluble for $n \geq 5$,⁷⁰ because any composition series for S_n must have A_n as one of its composition factors, and this is simple but nonabelian for $n \geq 5$.

If a group G is soluble, then so is any subgroup $H \leq G$,⁷¹ and any quotient G/H .⁷² Furthermore, extensions of soluble groups are soluble: if $H \triangleleft G$ and G/H are soluble, then so is G .⁷³ Consequently, finite direct products of soluble groups are also soluble,⁷⁴ as are finite abelian groups⁷⁵ and finite p -groups.⁷⁶

Two important results whose proofs are beyond the scope of this book are **Burnside's Theorem**,⁷⁷ which says that any group of order $p^m q^n$, for p and q both prime and $m, n \geq 0$, is soluble, and the **Feit–Thompson Theorem**,⁷⁸ which says that any group of odd order is soluble.

The **derived series** of a group G is defined recursively by using commutator subgroups. We set $G^{(0)} = G$, and let $G^{(i+1)}$ be the commutator $[G^{(i)}, G^{(i)}]$ for $i \geq 0$.⁷⁹ A group is soluble if and only if its derived series terminates at the trivial subgroup; that is, if $G^{(n)} = \{e\}$ for some $n \geq 0$.⁸⁰ The **derived length** of a soluble group G is the length of its derived series.⁸¹ The trivial group is the only group with derived length 0, the groups of derived length 1 are exactly the abelian groups,

⁶¹ Theorem 7.29, page 249.

⁶² Figure 7.2, page 249.

⁶³ Lemma 7.28, page 248.

⁶⁴ Definition 7.31, page 253.

⁶⁵ Proposition 7.32, page 253.

⁶⁶ Definition 7.33, page 253.

⁶⁷ Theorem 7.34, page 254.

⁶⁸ Proposition 7.35, page 255.

⁶⁹ Definition 7.36, page 255.

⁷⁰ Proposition 7.37, page 256.

⁷¹ Proposition 7.38, page 256.

⁷² Proposition 7.39, page 257.

⁷³ Proposition 7.40, page 257.

⁷⁴ Corollary 7.41, page 257.

⁷⁵ Proposition 7.42, page 258.

⁷⁶ Proposition 7.43, page 258.

⁷⁷ Theorem 7.44, page 258.

⁷⁸ Theorem 7.45, page 258.

⁷⁹ Definition 7.46, page 259.

⁸⁰ Proposition 7.47, page 259.

⁸¹ Definition 7.48, page 260.

and groups of derived length 2 are called **metabelian**.

Another important series is the **central series**, constructed using the centre. A normal series is said to be **central** if $G_{i+1}/G_i \leq Z(G/G_i)$ for all $0 \leq i < n$,⁸² that is, if each quotient G_{i+1}/G_i is contained in the centre of the corresponding quotient G/G_i of the full group. This construction only works with normal series, not subnormal ones, since we need each subgroup G_i to be normal in the full group G for the quotient G/G_i to be defined.

There are two special central series: the **upper central series**⁸³ and the **lower central series**.⁸⁴ The first of these is an ascending series, and is defined in terms of the **higher centres** $Z^i(G)$ of G : we set $G_0 = Z^0(G) = \{e\}$ and recursively define $Z^i(G)$ such that $Z^i(G)/Z^{i-1}(G) = Z(G/Z^{i-1}(G))$. The second is a descending series, and is defined by setting $G_0 = \Gamma_1(G) = G$ and recursively defining $G_i = \Gamma_{i+1}(G) = [G, \Gamma_i(G)]$.

The upper central series need not terminate at G , and the lower central series need not terminate at the trivial subgroup $\{e\}$. The lower central series is the optimal descending central series: if it reaches the trivial group, then it does so in fewer steps than any other descending central series.⁸⁵ Analogously, the upper central series is the optimal ascending central series: if it reaches G then it does so faster than any other ascending central series.⁸⁶

A group G with a finite-length central series connecting G with its trivial subgroup $\{e\}$ is said to be **nilpotent**, and the length of the shortest possible such central series is called the **nilpotency class** of G .⁸⁷ The nilpotency class of G will therefore be equal to the length of the upper or lower central series, whichever is smaller. A group G is therefore nilpotent if and only if its lower central series terminates at the trivial subgroup, and if and only if its upper central series terminates at G .⁸⁸

For any group G , the derived group $G^{(i)} \leq \Gamma_{i+1}(G)$ for all $i \geq 0$,⁸⁹ which means that all nilpotent groups are soluble.⁹⁰ Not all soluble groups are nilpotent, however: for example, the symmetric group S_3 is soluble because its derived series $\{1\} \triangleleft A_3 \triangleleft S_3$ reaches $\{1\}$ after finitely many steps. But its lower central series is $A_3 \triangleleft S_3$, which doesn't reach the trivial subgroup, so it isn't nilpotent.

Subgroups of nilpotent groups are also nilpotent,⁹¹ as are finite direct products of nilpotent groups,^{92,93} and so are homomorphic images and quotients of nilpotent groups,^{94,95} but extensions of nilpotent groups aren't in general nilpotent.

Finite p -groups are nilpotent,⁹⁶ and hence all finite direct products

⁸² Definition 7.49, page 261.

⁸³ Definition 7.51, page 262.

⁸⁴ Definition 7.53, page 263.

⁸⁵ Proposition 7.56, page 264.

⁸⁶ Proposition 7.58, page 265.

⁸⁷ Definition 7.55, page 264.

⁸⁸ Corollary 7.59, page 265.

⁸⁹ Proposition 7.60, page 266.

⁹⁰ Corollary 7.61, page 266.

⁹¹ Proposition 7.62, page 266.

⁹² Proposition 7.63, page 267.

⁹³ Corollary 7.64, page 267.

⁹⁴ Proposition 7.65, page 267.

⁹⁵ Corollary 7.66, page 268.

⁹⁶ Proposition 7.67, page 268.

of p -groups are nilpotent. In particular, any group that happens to be a direct product of its Sylow subgroups will be nilpotent. The converse also holds: a finite group G is nilpotent if and only if it is the direct product of its Sylow subgroups.⁹⁷ The proof of this result uses the fact that every proper subgroup H of a nilpotent group G is a proper subgroup of its normaliser $N_G(H)$.^{98,99} Also, a finite group G is nilpotent if and only if all its Sylow subgroups are normal.¹⁰⁰

Between the soluble and nilpotent groups there is another class of groups called **supersoluble** or **supersolvable**: groups which have normal series with cyclic quotients.¹⁰¹ Every supersoluble group is soluble, but not every soluble group is supersoluble: for example, S_4 and A_4 are soluble but not supersoluble since neither has a normal series with cyclic factors.¹⁰²

Subgroups of supersoluble groups are supersoluble,¹⁰³ as are quotients of supersoluble groups,¹⁰⁴ and finite direct products.¹⁰⁵

Extensions of supersoluble groups by supersoluble groups aren't necessarily supersoluble, but a slightly weaker result holds: extensions of supersoluble groups by cyclic groups are supersoluble.¹⁰⁶

We therefore have a hierarchy of classes of finite groups:

$$\text{trivial} \subset \text{cyclic} \subset \text{abelian} \subset \text{nilpotent} \subset \text{supersoluble} \subset \text{soluble}$$

If H and K are normal subgroups of some group G with trivial intersection $H \cap K = \{e\}$ and $HK = G$, then G is isomorphic to the direct product $H \times K$.^{107,108} Requiring only one of the subgroups, say H , to be normal, yields a more general construction: the **(internal) semidirect product**.¹⁰⁹ The subgroup K acts on the subgroup H by conjugation,¹¹⁰ so if a group G decomposes as an internal semidirect product HK then we automatically get a K -action on H . Conversely, given two groups H and K , together with a specified action of K on H , we can define a group structure on the Cartesian product $H \times K$, with $(h_1, k_1) * (h_2, k_2) = (h_1(k_1 \cdot h_2), k_1 k_2)$.¹¹¹ We call this group the **(external) semidirect product** of H and K , and denote it $H \rtimes K$.¹¹²

Many groups can be constructed as semidirect products of smaller ones. In particular, the dihedral group $D_n \cong \mathbb{Z}_n \rtimes \mathbb{Z}_2$,¹¹³ and the symmetric group $S_n \cong A_n \rtimes \mathbb{Z}_2$.¹¹⁴ The semidirect product $\mathbb{R}^n \rtimes GL_n(\mathbb{R})$ yields the **affine general linear group** $AGL_n(\mathbb{R})$.¹¹⁵

If we take the semidirect product of a group G with its full automorphism group $\text{Aut}(G)$ we obtain the **holomorph** of G , denoted $\text{Hol}(G) = G \rtimes \text{Aut}(G)$.¹¹⁶ Taking the semidirect product of n -fold direct product $G^n = G \times \cdots \times G$ with the symmetric group S_n , acting via permutation of coordinates, gives a group known as the **permutation wreath product** of G with S_n , denoted $G \wr S_n = G \text{ pwr } S_n = G^n \rtimes S_n$.

⁹⁷ Proposition 7.68, page 269.

⁹⁸ Proposition 7.69, page 269.

⁹⁹ Corollary 7.70, page 269.

¹⁰⁰ Corollary 7.72, page 270.

¹⁰¹ Definition 7.73, page 271.

¹⁰² Example 7.74, page 271.

¹⁰³ Proposition 7.75, page 271.

¹⁰⁴ Proposition 7.76, page 271.

¹⁰⁵ Proposition 7.77, page 271.

¹⁰⁶ Proposition 7.78, page 272.

¹⁰⁷ Proposition 2.12, page 46.

¹⁰⁸ Corollary 3.23, page 80.

¹⁰⁹ Definition 7.79, page 273.

¹¹⁰ Proposition 7.80, page 273.

¹¹¹ Proposition 7.81, page 274.

¹¹² Definition 7.82, page 275.

¹¹³ Example 7.83, page 275.

¹¹⁴ Example 7.84, page 275.

¹¹⁵ Example 7.85, page 276.

¹¹⁶ Example 7.86, page 276.

¹¹⁷ Theorem 2.13, page 47.

¹¹⁸ Example 7.87, page 277.

¹¹⁹ Definition 7.88, page 278.

¹²⁰ Definition 7.89, page 278.

¹²¹ Definition 7.90, page 279.

¹²² Definition 7.91, page 279.

¹²³ Proposition 7.92, page 280.

More generally, regarding some arbitrary group H as a permutation group on its underlying set, via Cayley's Theorem,¹¹⁷ we can form the **regular wreath product** of G by H , denoted $G \text{ wr } H$ (or $G \wr H$ if the meaning is clear from context). This construction enables us to define the **generalised symmetric group** $S_{m,n} = \mathbb{Z}_m \wr S_n$.¹¹⁸ In particular, $S_{1,n} = S_n$ is the usual symmetric group, $S_{2,n}$ is the **signed symmetric group** or **hyperoctahedral group**, and $S_{2,2} \cong D_4$.

Although many groups can be decomposed or constructed as semidirect products of smaller groups, there are also many that can't: the smallest example being $\mathbb{Z}_4$. A group G is an **extension** of a group K by a group H if there exists an inclusion homomorphism $i: H \hookrightarrow G$ with $i(H) \trianglelefteq G$, and $G/i(H) \cong K$.¹¹⁹ An inclusion function (not necessarily a homomorphism) $s: K \hookrightarrow G$ is called a **lifting** or a **section** if $s(e_K) = e_G$, and if $q \circ s = \text{id}_K$ where $q: G \rightarrow K$ is the quotient homomorphism.¹²⁰ Then $s(K)$ is copy of K injectively embedded as a subset, but not necessarily a subgroup, of G .

As with semidirect products, this section determines a conjugation K -action on H by $i(k \cdot h) = s(k)i(h)s(k)^{-1}$.

If s is a homomorphism, $G \cong H \rtimes K$ and we say the extension is **split**. Such a section is often called a **splitting** or **splitting homomorphism**.

Classifying group extensions is complicated in general, but the abelian case is relatively straightforward. We define a G -**module** to be an abelian group A equipped with a G -action,¹²¹ and we say that two extensions

$$A \xhookrightarrow{i_1} E_1 \xrightarrow{q_1} G \quad \text{and} \quad A \xhookrightarrow{i_2} E_2 \xrightarrow{q_2} G$$

are **equivalent** if there is an isomorphism $f: E_1 \rightarrow E_2$ such that $i_2 = f \circ i_1$ and $q_1 = q_2 \circ f$; that is, if the diagram

$$\begin{array}{ccccc} & & E_1 & & \\ & \nearrow i_1 & & \searrow q_1 & \\ A & & & & G \\ & \searrow i_2 & & \nearrow q_2 & \\ & & E_2 & & \end{array}$$

commutes.¹²² A **factor set** or **cocycle** is a function $\sigma: G \times G \rightarrow A$ such that

$$\sigma(e_G, g) = \sigma(g, e_G) = 0$$

for any $g \in G$, and

$$\sigma(g_1, g_2) + \sigma(g_1 g_2, g_3) = g_1 \cdot \sigma(g_2, g_3) + \sigma(g_1, g_2 g_3)$$

for all $g_1, g_2, g_3 \in G$.¹²³ A factor set is determined by a section of the extension, and different sections of the same extension yield factor sets that differ by a **coboundary** or **inner factor set**: a function $\tau: G \times G \rightarrow$

A such that

$$\tau(g, h) = \omega(g) - \omega(gh) + g \cdot \omega(h)$$

for all $g, h \in G$, and some function $\omega: G \rightarrow A$ satisfying $\omega(e_G) = 0$. Two factor sets related in this way are said to be **equivalent** or **cohomologous**.¹²⁴ Let $\text{FS}(G, A)$ be the set of factor sets of extensions of a group G by a G -module A , and let $\text{IFS}(G, A)$ be the corresponding subset of inner factor sets. These both form abelian groups under the canonical pointwise addition operation.¹²⁵ We define the group $\text{Ext}(G, A)$ to be the quotient $\text{FS}(G, A) / \text{IFS}(G, A)$; the elements of this group are equivalence classes of factor sets that differ only by an inner factor set. Furthermore, equivalent factor sets in this sense correspond to equivalent extensions,¹²⁶ so the group $\text{Ext}(G, A)$ classifies abelian extensions of G by A . The zero element in $\text{Ext}(G, A)$ corresponds to (the equivalence class of) the split extension $A \hookrightarrow A \rtimes G \twoheadrightarrow G$. The **Schur–Zassenhaus Theorem** says that if G is a group of order n and A is a G -module of order m , where m and n are coprime, then $\text{Ext}(G, A)$ is trivial; that is, the only extension of G by A is the split extension.¹²⁷

¹²⁴ Definition 7.95, page 283.

¹²⁵ Proposition 7.96, page 283.

¹²⁶ Proposition 7.97, page 284.

¹²⁷ Theorem 7.98, page 286.

An extension

$$A \xrightarrow{i} E \xrightarrow{q} G$$

is **central** if $i(A) \subseteq Z(E)$; that is, if the image of A lies in the centre of E , so that every element of $i(A)$ commutes with every element of E . We can, for example, construct the quaternion group Q_8 as a central extension of the Klein group V_4 by $\mathbb{Z}_2$.¹²⁸

¹²⁸ Example 7.99, page 287.

Another interesting case is that of a **cyclic extension**: an extension of a cyclic group G . Such an extension is determined entirely by a **cyclic extension type**, a quadruple (H, n, τ, v) where H is a (possibly nonabelian) group, $n \in \mathbb{N}$, $\tau \in \text{Aut}(H)$ and $v \in H$ such that $\tau(v) = v$ and τ^n is conjugation by v .¹²⁹ The **Cyclic Extension Theorem** says that for every cyclic extension type (H, n, τ, v) there is a group G with $H \trianglelefteq G$ and $G/H \cong \mathbb{Z}_n$; furthermore all extensions of $\mathbb{Z}_n$ are determined by some cyclic extension type.¹³⁰ Two cyclic extension types (H, n, τ, v) and (K, m, σ, w) are **equivalent** if $m = n$ and there exists an isomorphism $\phi: H \rightarrow K$ such that $\phi \circ \tau = \sigma \circ \phi$ and $\phi(v) = w$.¹³¹ Equivalent cyclic extension types yield isomorphic groups,¹³² although the converse doesn't necessarily hold: inequivalent cyclic extension types might realise isomorphic groups. Nevertheless, given two cyclic extension types (H, n, τ, v) and (H, n, σ, w) , if σ and τ are conjugate in $\text{Aut}(H)$ then they realise isomorphic groups.¹³³ We can use cyclic extension types to classify groups of certain orders.¹³⁴

¹²⁹ Definition 7.100, page 288.

¹³⁰ Theorem 7.101, page 288.

¹³¹ Definition 7.102, page 289.

¹³² Proposition 7.103, page 289.

¹³³ Proposition 7.104, page 290.

¹³⁴ Example 7.105, page 290.

Metabelian groups, those with derived length 2 can also be characterised as those groups that can be constructed as extensions of abelian groups by abelian groups.¹³⁵

¹³⁵ Definition 7.106, page 290.

$ G $	number	G
1	1	$\{e\}$
2	1	$\mathbb{Z}_2$
3	1	$\mathbb{Z}_3$
4	2	$\mathbb{Z}_4, V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$
5	1	$\mathbb{Z}_5$
6	2	$\mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3, D_3$
7	1	$\mathbb{Z}_7$
8	5	$\mathbb{Z}_8, \mathbb{Z}_2 \oplus \mathbb{Z}_4, \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2, D_4, Q_8$
9	2	$\mathbb{Z}_9, \mathbb{Z}_3 \oplus \mathbb{Z}_3$
10	2	$\mathbb{Z}_{10}, D_5$
11	1	$\mathbb{Z}_{11}$
12	5	$\mathbb{Z}_{12} \cong \mathbb{Z}_4 \oplus \mathbb{Z}_3, \mathbb{Z}_2 \oplus \mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3, D_6, Dic_3, A_4$
13	1	$\mathbb{Z}_{13}$
14	2	$\mathbb{Z}_{14}, D_7$
15	1	$\mathbb{Z}_{15} \cong \mathbb{Z}_3 \oplus \mathbb{Z}_5$
16	14	$\mathbb{Z}_{16}, \mathbb{Z}_2 \oplus \mathbb{Z}_8, \mathbb{Z}_4 \oplus \mathbb{Z}_4, \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2, D_8, Dic_4, SD_{16}, D_4 \times \mathbb{Z}_2, Q_8 \times \mathbb{Z}_2, \mathbb{Z}_8 \rtimes_{\phi_3} \mathbb{Z}_2, V_4 \rtimes \mathbb{Z}_4, Q_8 \rtimes \mathbb{Z}_2, \mathbb{Z}_4 \rtimes \mathbb{Z}_4$
17	1	$\mathbb{Z}_{17}$
18	5	$\mathbb{Z}_{18} \cong \mathbb{Z}_9 \oplus \mathbb{Z}_2, \mathbb{Z}_3 \oplus \mathbb{Z}_6 \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_2, D_9, \mathbb{Z}_3 \times D_3, (\mathbb{Z}_3 \oplus \mathbb{Z}_3) \rtimes \mathbb{Z}_2$
19	1	$\mathbb{Z}_{19}$
20	5	$\mathbb{Z}_{20} \cong \mathbb{Z}_4 \oplus \mathbb{Z}_5, \mathbb{Z}_2 \oplus \mathbb{Z}_{10} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_5, D_{10}, Dic_5, AGL_1(5)$
21	2	$\mathbb{Z}_{21} \cong \mathbb{Z}_7 \oplus \mathbb{Z}_3, \mathbb{Z}_7 \rtimes \mathbb{Z}_3$
22	2	$\mathbb{Z}_{22}, D_{11}$
23	1	$\mathbb{Z}_{23}$
24	15	$\mathbb{Z}_{24} \cong \mathbb{Z}_3 \oplus \mathbb{Z}_8, \mathbb{Z}_2 \oplus \mathbb{Z}_{12} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}_3, \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3, S_4, D_{12}, \mathbb{Z}_3 \times D_4, \mathbb{Z}_3 \times Q_8, \mathbb{Z}_2 \times A_4, \mathbb{Z}_4 \times D_3, \mathbb{Z}_2 \times Dic_3, \mathbb{Z}_2 \times D_6 \cong V_4 \times D_3, Q_8 \rtimes \mathbb{Z}_3 \cong SL_2(3), \mathbb{Z}_3 \rtimes \mathbb{Z}_8, \mathbb{Z}_3 \rtimes D_4, \mathbb{Z}_3 \rtimes Q_8$
25	2	$\mathbb{Z}_{25}, \mathbb{Z}_5 \oplus \mathbb{Z}_5$
26	2	$\mathbb{Z}_{26}, D_{13}$
27	5	$\mathbb{Z}_{27}, \mathbb{Z}_3 \oplus \mathbb{Z}_9, Heis(3) = U(3, 3), \mathbb{Z}_9 \rtimes \mathbb{Z}_3$
28	4	$\mathbb{Z}_{28} \cong \mathbb{Z}_4 \oplus \mathbb{Z}_7, \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_7 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_{14}, Dic_7, D_{14}$
29	1	$\mathbb{Z}_{29}$
30	4	$\mathbb{Z}_{30} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_5, D_{15}, D_3 \times \mathbb{Z}_5, D_5 \times \mathbb{Z}_3$
31	1	$\mathbb{Z}_{31}$

Table 7.19: Groups of order at most 31

We can use the results developed in this and earlier chapters, particularly Sylow's Theorems and the theory of semidirect products and group extensions, to classify groups of order less than 32. These are listed in Table 7.19.

References and further reading

Some interesting details about the history of Sylow's Theorems can be found in the following article:

- W C Waterhouse, *The early proofs of Sylow's theorem*, Archive for History of Exact Sciences 21.3 (1980) 279–290

In the spirit of James McKay's concise proof of Cauchy's Theorem, cited at the end of Chapter 2, Mícheál Ó Searcóid formulated shorter proofs of Sylow's Theorems:

- M Ó Searcóid, *A reordering of the Sylow Theorems*, The American Mathematical Monthly 94.2 (1987) 165–168

An accessible introduction to semidirect products, motivated by affine transformations (discussed in Example 7.85), can be found in the following article:

- S S Abhyankar and C Christensen, *Semidirect products: $x \mapsto ax+b$ as a first example*, Mathematics Magazine 75.4 (2002) 284–289

For more detailed discussion of homological algebra, the following books are good places to start:

- K S Brown, *Cohomology of Groups*, Graduate Texts in Mathematics 87, Springer (1994)
- P J Hilton and U Stammbach, *A Course in Homological Algebra*, second edition, Graduate Texts in Mathematics 4, Springer (1996)
- J J Rotman, *An Introduction to Homological Algebra*, first edition, Pure and Applied Mathematics 85, Academic Press (1979)
- J J Rotman, *An Introduction to Homological Algebra*, second edition, Universitext, Springer (2008)
- C A Weibel, *An Introduction to Homological Algebra*, Cambridge Studies in Advanced Mathematics 38, Cambridge University Press (1994)

All are aimed at graduate or advanced undergraduate students; Weibel and the second edition of Rotman are pitched at a higher level than the others. Brown's book focuses mostly on applications to group theory, while the others take a more general viewpoint concerning modules over arbitrary rings. As noted earlier, the classification of groups of order 16 in Section 7.A closely follows the exposition in the very clear and readable article by Marcel Wild:

- M Wild, *The groups of order sixteen made easy*, The American Mathematical Monthly 112 (2005) 20–31

Another classification of groups of order less than 32 can be found in the textbook by John Moody:

- J A Moody, *Groups for Undergraduates*, World Scientific (1994)

Exercises

8.1 Construct the dihedral group D_4 as a central extension of V_4 by $\mathbb{Z}_2$.

8 Rings

IN CHAPTER 1 we carefully studied the set $\mathbb{Z}$ of integers, giving particular attention to its additive structure. By deconstructing and abstracting the way addition of integers works, we derived the concept of a binary operation defined on an arbitrary set, and also a number of properties (associativity, commutativity, existence of inverses and an identity element) that such an operation might satisfy. This led us to formulate the definition of a group, a surprisingly rich and strange concept that has kept us busy for 317 pages so far.

But we can do other things with integers as well as add them together. The second arithmetical operation we learn as children is that of multiplication. In the next section we will study the multiplicative structure of $\mathbb{Z}$, both on its own and, crucially, how it interacts with the additive structure. This will lead us to define a new mathematical object, a **ring**, consisting of a set equipped with two binary operations.

8.1 Numbers

CONSIDERING JUST THE ADDITIVE STRUCTURE OF THE INTEGERS led us fairly naturally to the concept of an abelian group. We found shortly afterwards that many important examples of the structures we were interested in didn't satisfy the commutativity condition: in particular, matrix multiplication and the composition of symmetry operations or permutations led to interesting nonabelian groups.

However, at present we're primarily interested in algebraic structures that mimic the usual addition and multiplication operations on the integers, so for the moment we'll start with an additive abelian group.

So, what properties does integer multiplication satisfy? It's associative: for any integers a , b and c it's always the case that $(ab)c = a(bc)$. Associativity turns out to be important, so we'll put it on our list.

Integer multiplication is also commutative: $ab = ba$ for any integers a

Fafner: New mischief will the Nibelung plot against us if the gold gives him power. You there, Loge! Say without lies: of what great value is the gold then, that it satisfies the Nibelung?

Loge: It is a toy in the depths of the water, to give pleasure to laughing children; but if it were fashioned into a round ring it would bestow supreme power and win its master the world.

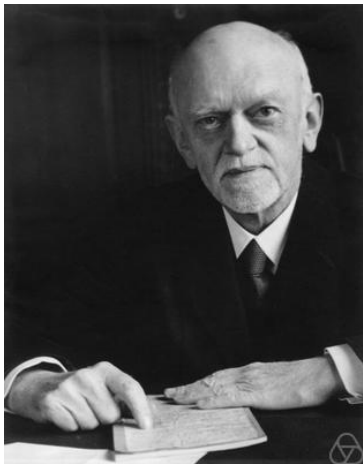
Wotan: I have heard talk of the Rhine's gold: its glittering glow hides runes of riches; a ring would give unbounded power and wealth.

— Richard Wagner (1813–1883),
Das Rheingold (1869)

"Can you do addition?" the White Queen asked. "What's one and one and one and one and one and one and one and one and one?"

"I don't know," said Alice. "I lost count."

— Lewis Carroll (Charles Lutwidge Dodgson) (1832–1898),
Alice's Adventures in Wonderland (1865)



Oberwolfach Photo Collection / L. Reidemeister

David Hilbert (1862–1943) was one of the most influential mathematicians of the late 19th and early 20th centuries.

Born in Königsberg, in East Prussia, he enrolled at the “Albertina” (the University of Königsberg) in 1880, obtaining his doctorate in 1885 for a thesis on spherical functions and binary forms, written under the supervision of Ferdinand von Lindemann (1852–1939).

In 1888 he proved the celebrated *Hilbert Basis Theorem*, an innovative solution to a problem set twenty years earlier by Paul Gordan (1837–1912), who is said to have remarked “Das ist nicht Mathematik, das ist Theologie” (“This is not mathematics, this is theology.”)

He stayed at Königsberg until 1895, when he was appointed to a chair at the University of Göttingen. In addition to his many contributions to geometry, physics, analysis and algebra, he fostered a vibrant academic environment at Göttingen, which succeeded in attracting active research mathematicians from around the world.

At the International Congress of Mathematicians in Paris, in 1900, he presented a famous list of ten important unsolved problems, later extended to twenty-three, of which four remain unresolved over a century later.

He died in 1943 and was buried in Göttingen. His epitaph consists of a remark made at the end of his retirement address in 1930 to the Society of German Scientists and Physicians: “Wir müssen wissen. Wir werden wissen.” (“We must know. We will know.”)

and b . However, when we investigated groups, we found that lots of important and interesting examples didn’t satisfy the commutativity requirement, so although integer multiplication does, we might want to put this on the optional list.

There is also a special integer 1 that acts as a multiplicative identity element, in the sense that $a1 = 1a = a$ for any $a \in \mathbb{Z}$. Notice that $1 \neq 0$; that is, the multiplicative identity isn’t the same as the additive identity. We’ll see in a little while that this is (almost) always true. Something else that’s worth mentioning at this stage is that although $\mathbb{Z}$ and many of the other structures of this type we’re going to study do have multiplicative identities, we’ll also meet some examples which don’t. So we’ll put this one on the optional list too. This also leads us into a slight ambiguity of terminology that we’ll deal with soon.

What about multiplicative inverses? Well, $\mathbb{Z}$ doesn’t have them in general. In fact, the only invertible elements in $\mathbb{Z}$ are 1 and -1 . Later on, we’ll consider structures which do have invertible elements, but for the moment it’s not something we’re going to require.

That’s pretty much it for the multiplicative operation on its own, but we also need to think about how multiplication and addition interact with each other. There are really just two questions we need to ask here: what happens if we add two integers together and multiply on the left, or the right by another integer. That is, how do we expand expressions of the form $a(b + c)$ and $(a + b)c$?

We know how to do this for integers: we all learned it quite early on in our mathematical education. The expressions we want are, in general,

$$a(b + c) = ab + ac, \quad (a + b)c = ac + bc.$$

We call these the (left and right) **distributive laws**.

We’re now ready to write down a formal definition of the objects we’re going to be studying for most of the rest of the book.

Definition 8.1 A **ring** $R = (R, +, \cdot)$ is a set R equipped with two binary operations $+$ and $\cdot$ such that:

- R1** $(R, +)$ is an abelian group;
- R2** $\cdot$ is associative, that is, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in R$;
- R3** $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$ for all $a, b, c \in R$;
- R4** there exists a **unity** or **multiplicative identity** element $1 \in R$ such that $1 \cdot a = a \cdot 1 = a$ for all $a \in R$.

A **commutative ring** is one that also satisfies the following extra condition:

- R5** $a \cdot b = b \cdot a$ for all $a, b \in R$.

Axiom R2 says that $(R, \cdot)$ is a semigroup, while R4 says that it is a monoid. If R5 is satisfied as well, then $(R, \cdot)$ is a commutative monoid. The above definition says that a ring is an abelian group equipped with another binary operation under which it forms a monoid, and that both operations are compatible with each other in a specific way.

Now to attend to the terminological issue mentioned earlier. Some books use the term **ring** to refer to a slightly more general structure that doesn't necessarily satisfy condition R4, and call the structure in Definition 8.1 a **ring with 1**, a **unital ring** or a **ring with unity**.

Those books (such as this one) that use the term **ring** for the corresponding structure that does have a multiplicative identity sometimes refer to the more general case as a **ring without 1**, a **nonunital ring**, a **pseudoring** or, most succinctly but least pronouncably, as a **rng**.

It's time to look at some concrete examples.

Example 8.2 The set $\mathbb{Z}$ of integers forms a commutative ring with respect to the familiar addition and multiplication operations.

Similarly, the set $\mathbb{Q}$ of rational numbers, the set $\mathbb{R}$ of real numbers and the set $\mathbb{C}$ of complex numbers all form commutative rings under the usual addition and multiplication operations.

The rings $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ have additional properties that the integers don't, and we'll study these in more depth over the next few chapters.

The next examples of groups we met in Chapter 1 were the finite cyclic groups $\mathbb{Z}_n$. These are abelian groups, and with a little extra work we can turn them into rings.

Example 8.3 The finite cyclic group $\mathbb{Z}_n = \{0, \dots, n-1\}$ forms a commutative ring when equipped with modulo- n multiplication.

To see this, we check the five conditions from Definition 8.1. The pair $(\mathbb{Z}_n, +_n)$ is an abelian group, so R1 is satisfied. Modulo- n multiplication is associative, commutative, and also distributive with respect to modulo- n addition, so conditions R2, R3 and R5 hold. The element $1 \in \mathbb{Z}_n$ behaves as a multiplicative identity, so condition R4 is satisfied too, and hence $\mathbb{Z}_n$ is a commutative ring. The addition and multiplication tables for $\mathbb{Z}_2$, $\mathbb{Z}_3$ and $\mathbb{Z}_4$ are shown in Table 8.1.

The simplest example of a ring is the set consisting of a single element:

Example 8.4 Let $R = \{0\}$. This forms a commutative unital ring, the **trivial ring**, under the usual addition and multiplication operations

$$0 + 0 = 0 \quad \text{and} \quad 0 \cdot 0 = 0.$$

$+$	0	1
0	0	1
1	1	0

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

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

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

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

Table 8.1: Addition and multiplication tables for $\mathbb{Z}_2$, $\mathbb{Z}_3$ and $\mathbb{Z}_4$

The following two examples are formed by taking the ring $\mathbb{Z}$ and carefully attaching a specific non-integer element. We'll come back to

this idea later on when we study field extensions, but for the moment it's an interesting and useful way of constructing new rings.

Example 8.5 The set $\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$ is a commutative unital ring under the usual addition and multiplication operations. It is an abelian group under addition, so R1 holds. Multiplication is associative and commutative, and also distributive over addition, so R2, R3 and R5 hold. Finally, $1 = 1 + 0\sqrt{2} \in \mathbb{Z}[\sqrt{2}]$ so R4 holds too.

Example 8.6 A variation on the above example is the ring

$$\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$$

of **Gaussian integers**. This is a commutative unital ring.

Similarly, the set of **Eisenstein integers**

$$\mathbb{Z}[\omega] = \{a + b\omega : a, b \in \mathbb{Z}\},$$

where $\omega = \frac{1}{2} + \frac{\sqrt{3}}{2}i$, and the set of **Kleinian integers**

$$\mathbb{Z}[\lambda] = \{a + b\lambda : a, b \in \mathbb{Z}\},$$

where $\lambda = \frac{1}{2} + \frac{\sqrt{2}}{2}i$, also both form commutative rings.

We can modify the above examples by replacing $\mathbb{Z}$ with some other ring like $\mathbb{Q}$ or $\mathbb{Z}_n$, to get new rings such as $\mathbb{Q}[i]$ or $\mathbb{Z}_n[\sqrt{2}]$.

Example 8.7 Let $\omega = \frac{1}{2} + \frac{\sqrt{3}}{2}i$. Then

$$\mathbb{Z}_2[i] = \{a + bi : a, b \in \mathbb{Z}_2\},$$

$$\mathbb{Z}_2[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Z}_2\},$$

$$\mathbb{Z}_2[\omega] = \{a + b\omega : a, b \in \mathbb{Z}_2\}.$$

The addition and multiplication tables for these rings are shown in Tables 8.2 and 8.3. (Note that $\omega + 1 = -\omega^2$, but since we're working over $\mathbb{Z}_2$ we have $+1 = -1$ and hence $\omega + 1 = \omega^2$.)

Now we've met a few examples of rings, it's time to start formulating some general results about them. The following basic properties are almost immediate consequences of Definition 8.1.

Proposition 8.8 Let $R = (R, +, \cdot)$ be a (unital or nonunital) ring. Then, for any $a, b, c \in R$,

- (i) $0 \cdot a = a \cdot 0 = 0$,
- (ii) $a \cdot (-b) = (-a) \cdot b = -(a \cdot b)$, and
- (iii) $(-a) \cdot (-b) = -(a \cdot b)$.

If R is unital, then

- (iv) $(-1) \cdot a = -a = a \cdot (-1)$ for any $a \in R$, and
- (v) $(-1) \cdot (-1) = 1$.

+	0	1	1+i	i
0	0	1	1+i	i
1	1	0	i	1+i
1+i	1+i	i	0	1
i	i	1+i	1	0

+	0	1	$\sqrt{2}$	$1+\sqrt{2}$
0	0	1	$\sqrt{2}$	$1+\sqrt{2}$
1	1	0	$1+\sqrt{2}$	$\sqrt{2}$
$\sqrt{2}$	$\sqrt{2}$	$1+\sqrt{2}$	0	1
$1+\sqrt{2}$	$1+\sqrt{2}$	$\sqrt{2}$	1	0

+	0	1	ω	ω^2
0	0	1	ω	ω^2
1	1	0	ω^2	ω
ω	ω	ω^2	0	1
ω^2	ω^2	ω	1	0

Table 8.2: Addition tables for the rings $\mathbb{Z}_2[i]$, $\mathbb{Z}_2[\sqrt{2}]$ and $\mathbb{Z}_2[\omega]$

$\cdot$	0	1	1+i	i
0	0	0	0	0
1	0	1	1+i	i
1+i	0	1+i	0	1+i
i	0	i	1+i	1

$\cdot$	0	1	$\sqrt{2}$	$1+\sqrt{2}$
0	0	0	0	0
1	0	1	$\sqrt{2}$	$1+\sqrt{2}$
$\sqrt{2}$	0	$\sqrt{2}$	0	$\sqrt{2}$
$1+\sqrt{2}$	0	$1+\sqrt{2}$	$\sqrt{2}$	1

$\cdot$	0	1	ω	ω^2
0	0	0	0	0
1	0	1	ω	ω^2
ω	0	ω	ω^2	1
ω^2	0	ω^2	1	ω

Table 8.3: Multiplication tables for the rings $\mathbb{Z}_2[i]$, $\mathbb{Z}_2[\sqrt{2}]$ and $\mathbb{Z}_2[\omega]$

Proof All of these follow from the ring axioms R1, R2 and R3, and R5 in the case of (iv) and (v). For (i),

$$0 + a \cdot 0 = a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0$$

and using the cancellation law for abelian groups we get $0 = a \cdot 0$ as required. The proof that $0 \cdot a = 0$ is very similar.

We need to think a little carefully about part (ii) and properly understand what it says. In particular, $-(a \cdot b)$ is the additive inverse of $a \cdot b$; that is, the unique (by Proposition 1.16) element of R that yields 0 when added to $a \cdot b$. So, to show that $(-a) \cdot b = -(a \cdot b)$ we need to show that $(-a) \cdot b + a \cdot b = 0$.

By the right distributive law,

$$(-a) \cdot b + a \cdot b = (-a + a) \cdot b = 0 \cdot b$$

and by part (i) we know that $0 \cdot b = 0$, hence $(-a) \cdot b = -(a \cdot b)$ as claimed. A very similar argument, using the left distributive law, shows that $a \cdot (-b) = -(a \cdot b)$.

We can use part (ii) to prove part (iii): note that $(-a) \cdot (-b) = -(a \cdot (-b))$, and furthermore that $-(a \cdot (-b)) = -(-(a \cdot b))$. This element $-(-(a \cdot b))$ is the unique element of R that yields 0 when added to $-(a \cdot b)$. But $a \cdot b$ satisfies this property, and by Proposition 1.16 we know that inverses are unique, so $-(-(a \cdot b)) = a \cdot b$, and hence $(-a) \cdot (-b) = a \cdot b$.

To show part (v) we use (ii) again: $(-1) \cdot a = -(1 \cdot a) = -a$ and $a \cdot (-1) = -(a \cdot 1) = -a$.

Finally, part (v) is the special case of (iii) where $a = b = 1$. $\square$

The above proof relies in a couple of places on the uniqueness of additive identities and inverses guaranteed by Proposition 1.16. This works because a ring is essentially an abelian group with some extra structure bolted onto it. It's worth asking whether the multiplicative identity is also unique, and it turns out that it is.

Proposition 8.9 Let $R = (R, +, \cdot)$ be a ring with a multiplicative identity 1. This identity element is unique.

Proof Suppose that there is another element $e \in R$ such that $e \cdot a = a \cdot e = a$ for all $a \in R$. Then $e \cdot 1 = 1$, because e is a multiplicative identity. But also $e \cdot 1 = e$ since 1 is a multiplicative identity too. Equating these gives $1 = e \cdot 1 = e$, so the identity must be unique. $\square$

All of the examples we've seen so far have been commutative and unital, so it's time we saw some that aren't. The following is a simple example of a commutative nonunital ring.



Wikimedia Commons / Christian Albrecht Jensen (1792–1870)

Carl Friedrich Gauss (1777–1855) is widely regarded as the greatest mathematician of the late 18th and early 19th centuries. Born to poor, working-class parents in Braunschweig (Brunswick), he displayed considerable mathematical talents from early childhood. Between 1792 and 1799, during his undergraduate studies at the Collegium Carolinum in Braunschweig and the University of Göttingen, he independently discovered several important mathematical and physical results, including Bode's Law, the Binomial Theorem, the Quadratic Reciprocity Law and the Prime Number Theorem. Also during this period he wrote *Disquisitiones Arithmeticae*, his celebrated treatise on number theory, which was published a few years later in 1801.

Under the supervision of Johann Friedrich Pfaff (1765–1825), he obtained a doctorate from the University of Helmstadt in 1799 for a thesis on the Fundamental Theorem of Algebra.

His work in differential geometry includes the *Theorema Egregium* and the Gauss–Bonnet Theorem, and his unpublished research on non-Euclidean geometry anticipated later work by János Bolyai (1802–1860) and Nikolai Lobachevsky (1792–1856).

In addition to his mathematical achievements, he made important contributions in astronomy. In particular, in late 1801 he successfully predicted the location of the dwarf planet Ceres, which had been discovered by the Italian astronomer Giuseppe Piazzi (1746–1826) earlier that year and then lost as it passed behind the sun.

Example 8.10 The set $2\mathbb{Z} = \{\dots, -4, -2, 0, 2, 4, \dots\}$ of even integers forms a commutative nonunital ring under the usual addition and multiplication operations.

The pair $(2\mathbb{Z}, +)$ is an abelian group, so R_1 holds. Multiplication is associative, distributive and commutative so R_2 , R_3 and R_5 are satisfied. The product of two even integers is also an even integer, so multiplication is well-defined on $2\mathbb{Z}$. But the identity axiom R_4 fails: there is no even integer n such that $a \cdot n = n \cdot a = a$ for all $a \in 2\mathbb{Z}$.

This example generalises: for some fixed $n \in \mathbb{Z}$ let

$$n\mathbb{Z} = \{\dots, -2n, -n, 0, n, 2n, \dots\}.$$

Then for $|n| > 1$ the above argument shows that $n\mathbb{Z}$ is a commutative nonunital ring. If $|n| = 1$ then $n\mathbb{Z} = \mathbb{Z}$, which is a commutative unital ring as seen in Example 8.2. If $n = 0$ then we get the trivial ring $\{0\}$.

Another source of nonunital rings is shown in the next example, where we take an abelian group and give it a trivial multiplication operation.

Example 8.11 Let A be an abelian group with identity e . Define $a \cdot b = e$ for all $a, b \in A$. This process forms a nonunital ring.

In the case of the group $\mathbb{Z}_4$, this process yields a ring with the following addition and multiplication tables:

$+$	0	1	2	3	$\cdot$	0	1	2	3
0	0	1	2	3	0	0	0	0	0
1	1	2	3	0	1	0	0	0	0
2	2	3	0	1	2	0	0	0	0
3	3	0	1	2	3	0	0	0	0

We'll meet an important class of noncommutative rings in the next section, but we can look at two interesting examples now.

Example 8.12 The sets

$$\mathbb{H} = \{a + bi + cj + dk : a, b, c, d \in \mathbb{R}\}$$

$$\text{and } \mathbb{L} = \{a + bi + cj + dk : a, b, c, d \in \mathbb{Z}\}$$

of, respectively, **quaternions** and **Lipschitz integers** (or **Lipschitz quaternions**) both form noncommutative unital rings under the usual quaternion addition and multiplication operations, where $i^2 = j^2 = k^2 = -1$ and

$$\begin{aligned} i \cdot j &= k, & j \cdot k &= i, & k \cdot i &= j, \\ j \cdot i &= -k, & k \cdot j &= -i, & i \cdot k &= -j. \end{aligned}$$

In both cases the conditions R_1 , R_2 and R_3 are satisfied and R_4 holds because $1 = 1 + 0i + 0j + 0k$, but R_5 fails.

In all of the nontrivial examples of unital rings we've seen so far, the multiplicative identity element is different to the additive identity. The following proposition confirms this in general.

Proposition 8.13 *Let R be a unital ring in which $1 = 0$; that is, the additive and multiplicative identities are the same. Then R is trivial.*

Proof Suppose that $1 = 0$. Then for any $a \in R$ we have

$$a = 1 \cdot a = 0 \cdot a = 0$$

and hence R is trivial. □

The rings in Example 8.7 all have addition tables structurally identical to that of the Klein group $V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$. This leads us to ask whether any of them are structurally identical *as rings* to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$. But to answer this question we first need to decide what it means for two rings to be isomorphic, and also what we mean by the direct sum of two rings.

For the first one, recall that a group isomorphism is a bijection that respects the group structure in a precisely-defined way.² Such a bijection must respect the additive structure of an abelian group, and since a ring is to some extent just a special kind of abelian group, this is a good place to start. But a ring also has a multiplicative structure, so a ring isomorphism should respect that too.

Rather than define ring isomorphisms and then introduce homomorphisms as the non-bijective case (as we did with groups) we might as well introduce both concepts at the same time, although we'll leave the main discussion of ring homomorphisms until Chapter 9.

Definition 8.14 Let $R = (R, +, \cdot)$ and $S = (S, \boxplus, \boxdot)$ be rings, and suppose that $f: R \rightarrow S$ is a function satisfying

$$f(a + b) = f(a) \boxplus f(b) \quad \text{and} \quad f(a \cdot b) = f(a) \boxdot f(b)$$

for all $a, b \in R$. Then f is a **(ring) homomorphism**.

If f is injective, we call it a **monomorphism**; if f is surjective, we call it a **epimorphism**; and if f is bijective, we call it a **isomorphism**. A homomorphism $f: R \rightarrow R$ is called an **endomorphism**.

The second item on our list is the direct sum of two rings. Again, we define this by extending the corresponding concept for abelian groups:

Definition 8.15 Let R and S be rings. The **direct sum** $R \oplus S$ is

$$R \oplus S = \{(r, s) : r \in R, s \in S\}$$

with addition and multiplication operations

$$(a, b) + (c, d) = (a + c, b + d) \quad (a, b) \cdot (c, d) = (a \cdot c, b \cdot d)$$

for all $a, c \in R$ and $b, d \in S$.

There is a popular story that an eminent mathematician once spent several hundred pages formally proving that $1+1=2$. Unlike many such apocryphal legends, this one is true. The proof occurs on page 379 of the first volume of *Principia Mathematica*, Bertrand Russell and Alfred North Whitehead's notationally dense three-volume work on the axiomatic foundations of mathematics.¹

Here we have taken only 323 pages to confirm that $1 \neq 0$.

¹ B A W Russell and A N Whitehead, *Principia Mathematica*, Cambridge University Press (1910–1913)

² Definition 1.17, page 11.

+	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	(0,1)	(0,0)	(1,1)	(1,0)
(1,0)	(1,0)	(1,1)	(0,0)	(0,1)
(1,1)	(1,1)	(1,0)	(0,1)	(0,0)

·	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,0)	(0,0)	(0,0)
(0,1)	(0,0)	(0,1)	(0,0)	(0,1)
(1,0)	(0,0)	(0,0)	(1,0)	(1,0)
(1,1)	(0,0)	(0,0)	(1,0)	(1,1)

Table 8.4: Addition and multiplication tables for the ring $\mathbb{Z}_2 \oplus \mathbb{Z}_2$

So we're now able to ask whether or not any of the rings in Example 8.7 are isomorphic *as rings* to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

The addition and multiplication tables for $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ are shown in Table 8.4. This ring and those in Example 8.7 are all small, finite rings, so in principle we could check for an isomorphism by considering all possible functions from $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ to the ring in question. But this is not a particularly illuminating, clever or efficient method, and it rapidly becomes impractical for larger rings, and impossible for infinite rings. So we need to use more sophisticated methods.

However, both $\mathbb{Z}_2[i]$ and $\mathbb{Z}_2[\sqrt{2}]$ have a nontrivial element (respectively, $1+i$ and $\sqrt{2}$) that squares to give 0, but $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ has no such element, so neither $\mathbb{Z}_2[i]$ nor $\mathbb{Z}_2[\sqrt{2}]$ can be isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

Furthermore, $\mathbb{Z}_2[\omega]$ has a nontrivial element ω that cubes to give 1, and $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ doesn't. Nor, for that matter, does either $\mathbb{Z}_2[i]$ or $\mathbb{Z}_2[\sqrt{2}]$.

So, we know now that $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ isn't isomorphic to any of the rings $\mathbb{Z}_2[i]$, $\mathbb{Z}_2[\sqrt{2}]$ or $\mathbb{Z}_2[\omega]$. We also know that $\mathbb{Z}_2[\omega]$ isn't isomorphic to either $\mathbb{Z}_2[i]$ or $\mathbb{Z}_2[\sqrt{2}]$.

Are $\mathbb{Z}_2[i]$ and $\mathbb{Z}_2[\sqrt{2}]$ isomorphic? Well, a careful examination of their multiplication tables suggests they might be: both have 0 as an additive identity and 1 as a multiplicative identity, so we might reasonably expect an isomorphism $f: \mathbb{Z}_2[i] \rightarrow \mathbb{Z}_2[\sqrt{2}]$ to give $f(0) = 0$ and $f(1) = 1$. In $\mathbb{Z}_2[i]$ the element i squares to give 1, and there is only one remaining element of $\mathbb{Z}_2[\sqrt{2}]$ that squares to $f(1) = 1$, namely $1+\sqrt{2}$. Similarly, $(1+i)^2 = 0$ in $\mathbb{Z}_2[i]$, and $\sqrt{2}$ is the only remaining element of $\mathbb{Z}_2[\sqrt{2}]$ that squares to $f(0) = 0$. We have thus explicitly constructed an isomorphism $f: \mathbb{Z}_2[i] \rightarrow \mathbb{Z}_2[\sqrt{2}]$:

$$f(0) = 0, \quad f(1) = 1, \quad f(i) = 1+\sqrt{2}, \quad f(1+i) = \sqrt{2}.$$

Here we made use of the multiplicative structure of the rings in question, because a ring isomorphism must preserve this structure. In particular, we used the following facts about ring homomorphisms:

Proposition 8.16 *Let R and S be rings, and $f: R \rightarrow S$ be a ring homomorphism. Then $f(0) = 0$. Furthermore, if R and S both have a multiplicative identity 1, S is nontrivial, and f is surjective, then $f(1) = 1$. Also, $f(nr) = nf(r)$ and $f(r^n) = f(r)^n$ for any $r \in R$ and $n \in \mathbb{N}$.*

Proof The first statement follows from Proposition 4.24, since any ring R is an abelian group, and any ring homomorphism f satisfies the usual conditions for an abelian group homomorphism.

For the second statement, we want to show that $f(1) \cdot s = s \cdot f(1) = s$ for any $s \in S$. Since f is surjective, we can find at least one $r \in R$ such

that $f(r) = s$, and hence

$$f(1) \cdot s = f(1) \cdot f(r) = f(1 \cdot r) = f(r) = s.$$

(A very similar argument shows that $s \cdot f(1) = s$.)

The last statement follows from the definition of a ring homomorphism, and can be proved by induction on n . For $n = 1$ we have $f(1r) = f(r) = 1f(r)$ and $f(r^1) = f(r) = f(r)^1$.

Suppose that the hypothesis holds for $n = k$. That is, $f(kr) = kf(r)$ and $f(r^k) = f(r)^k$. Then

$$f((k+1)r) = f(kr + r) = f(kr) + f(r) = kf(r) + f(r) = (k+1)f(r)$$

and

$$f(r^{k+1}) = f(r^k \cdot r) = f(r^k) \cdot f(r) = f(r)^k \cdot f(r) = f(r)^{k+1}$$

as required. $\square$

Early on in our study of groups, we found it useful to consider the **order** $|g|$ of a group element g . We can extend this idea to elements of a ring, but since we have two binary operations to work with, we end up with two related types of order.

Definition 8.17 Let R be a ring, and $r \in R$ be some arbitrary element. The **additive order** of r is the smallest natural number n such that $nr = 0$; if no such n exists, the element r is said to have infinite additive order.

If R is a unital ring, then the **multiplicative order** of r is the smallest natural number n such that $r^n = 1$; again if no such n exists, we say r has infinite multiplicative order.

Unless otherwise specified, or where the context is ambiguous, we will typically use the term **order** to refer to the multiplicative order rather than the additive order.

As before, the **order** $|R|$ of a ring R is its cardinality. It will sometimes be useful to know the maximum additive order of elements in R :

Definition 8.18 The **characteristic** $\text{char}(R)$ of a ring R is the smallest natural number n such that $nr = 0$ for all $r \in R$. If no such n exists, we set $\text{char}(R) = 0$.

Over the next couple of chapters we will derive a number of important facts about the characteristic of various types of rings, but for the moment we can state and prove the following result about the characteristic of unital rings.

Proposition 8.19 Let R be a unital ring. If the multiplicative identity 1 has infinite additive order in R , then $\text{char}(R) = 0$; otherwise $\text{char}(R)$ is equal to the additive order of 1.

R	$ R $	$\text{char}(R)$	R	$ R $	$\text{char}(R)$
$\mathbb{Z}$	$\aleph_0$	0	$\mathbb{Z}_n$	n	n
$\mathbb{Q}$	$\aleph_0$	0	$2\mathbb{Z}$	$\aleph_0$	0
$\mathbb{R}$	$\aleph_1$	0	$\mathbb{Z}_2[i]$	4	2
$\mathbb{C}$	$\aleph_1$	0	$\mathbb{Z}_2[\omega]$	4	2
$\mathbb{H}$	$\aleph_1$	0	$\mathbb{Z}_2 \oplus \mathbb{Z}_2$	4	2

Table 8.5: Orders and characteristics of some rings

Proof Clearly if 1 has infinite additive order, then there is no $n \in \mathbb{N}$ such that $n1 = 0$, so by Definition 8.18, $\text{char}(R) = 0$.

If 1 has additive order n , then $nr = (n1) \cdot r = 0 \cdot r = 0$ for any $r \in R$, and n is the smallest natural number with this property. $\square$

Table 8.5 lists the characteristics of most of the rings we've met so far. We'll end this section with a result sometimes called the Freshman's, Schoolboy's or Child's Binomial Theorem.

Proposition 8.20 *Let R be a commutative ring of prime characteristic p . Then for any $a, b \in R$,*

$$(a + b)^p = a^p + b^p.$$

Proof By the usual binomial theorem, the coefficient of the term $a^n b^{n-p}$ is $\binom{p}{n} = \frac{p!}{n!(p-n)!}$. When $0 < n < p$ this coefficient is divisible by p , and hence is zero in R . Thus

$$(a + b)^p = a^p + 0a^{p-1}b + \cdots + 0ab^{p-1} + b^p = a^p + b^p$$

as claimed. $\square$

For this purpose we must commence, not with a square, but with an oblong arrangement of terms consisting, suppose, of m lines and n columns. This will not in itself represent a determinant, but is, as it were, a Matrix out of which we may form various systems of determinants by fixing upon a number p , and selecting at will p lines and p columns, the squares corresponding to which may be termed determinants of the p th order.

— James Joseph Sylvester (1814–1897),
Additions to the articles
"On a new Class of Theorems"
and "On Pascal's Theorems",
 Philosophical Magazine 37 (1850)
 363–370

8.2 Matrices

SOME OF THE FIRST EXAMPLES of nonabelian groups we met were the matrix groups $GL_n(\mathbb{R})$ and its various subgroups such as $O_n(\mathbb{R})$, $SO_n(\mathbb{R})$ and $SL_n(\mathbb{R})$, and it seems natural to ask whether this might be a source of interesting noncommutative rings.

Considering square matrices over some ring R , it's certainly the case that matrix addition is commutative and matrix multiplication is associative, so by carefully choosing a set of matrices to ensure closure under addition and multiplication, existence of additive identity and inverses, and so forth, we should be able to construct rings of matrices.

Example 8.21 Let R be a ring, and denote by $M_n(R)$ the set of $n \times n$ matrices with entries from R . Then $M_n(R)$ is a ring.

In particular, $M_2(\mathbb{Z}_3)$ is the ring of 2×2 matrices with entries from $\mathbb{Z}_3$. It has order $3^4 = 81$ and characteristic 3.

If we allow R to be nonunital, then $M_n(R)$ is also a nonunital ring.

Thus $M_3(2\mathbb{Z})$ is the nonunital ring of 3×3 matrices with even integer entries. It has (countably) infinite order and characteristic zero.

The ring $M_n(R)$ is commutative if $n = 1$ and R is itself commutative: in this case $M_1(R) \cong R$ via the obvious map $[r] \mapsto r$. However, $M_n(R)$ is in general noncommutative for $n > 1$.

In addition to the full matrix rings $M_n(R)$, some important subsets also form rings.

Example 8.22 Let R be a ring, and denote by $UT_n(R)$ the set of $n \times n$ **upper triangular matrices**; that is, $n \times n$ matrices which have all zero entries below the leading diagonal. More formally,

$$UT_n(R) = \{A \in M_n(R) : A_{ij} = 0 \text{ for } i < j\}.$$

Similarly,

$$LT_n(R) = \{A \in M_n(R) : A_{ij} = 0 \text{ for } i > j\}$$

is the set of **lower triangular matrices**, for which all the entries above the leading diagonal are zero.

Both of these sets form rings under the usual matrix addition and multiplication operations.

Finally, let

$$D_n(R) = \{A \in M_n(R) : A_{ij} = 0 \text{ for } i \neq j\}$$

be the set of **diagonal matrices**; this also forms a ring.

In fact, the ring $D_n(R)$ of diagonal $n \times n$ matrices over R is isomorphic to the n -fold direct sum $R^n = R \oplus \cdots \oplus R$, via the map

$$\begin{bmatrix} r_1 & 0 & 0 & \cdots & 0 \\ 0 & r_2 & 0 & \cdots & 0 \\ 0 & 0 & r_3 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & r_n \end{bmatrix} \mapsto (r_1, r_2, r_3, \dots, r_n).$$

From linear algebra we're used to using matrices to represent linear maps from one vector space to another. In particular, an $n \times n$ real matrix represents a linear map from $\mathbb{R}^n \rightarrow \mathbb{R}^n$. More generally, an element of $M_n(R)$ (or, for that matter, $UT_n(R)$, $LT_n(R)$ or $D_n(R)$) can be regarded as homomorphisms from R^n to R^n . Recall from Definition 8.14 that a ring homomorphism from a ring to itself is called an endomorphism, so every element of $M_n(R)$ is really just an endomorphism of R^n .

In particular, a given 1×1 matrix over R corresponds to an endomorphism of R . The way this works is that a matrix $[r] \in M_1(R)$ acts on an element $s \in R$ by multiplication on the left: $[r]s = r \cdot s$. This determines a unique endomorphism f_r of R given by $f_r(s) = r \cdot s$.

More generally, given any two endomorphisms f_1 and f_2 of R , we can define two new endomorphisms $(f_1 + f_2)$ and $(f_1 \cdot f_2)$ by

$$(f_1 + f_2)(r) := f_1(r) + f_2(r) \quad \text{and} \quad (f_1 \cdot f_2)(r) := f_1(f_2(r)) = (f_1 \circ f_2)(r)$$

for all $r \in R$. This leads to the following important example.

Example 8.23 Let $\text{End}(R) = \{f: R \rightarrow R\}$ be the set of all endomorphisms of a ring R . Then this set forms a ring, the **endomorphism ring** of R under the addition and multiplication operations defined above.

The additive identity of this ring is the zero endomorphism $z: R \rightarrow R$ mapping everything to the zero element $0 \in R$, and for any endomorphism f there is a corresponding endomorphism $-f$ given by $(-f)(r) = -(f(r))$ for all $r \in R$. Addition of endomorphisms is clearly associative, because addition in R is. Multiplication is associative too, because composition of endomorphisms is. Finally, the left and right distributive laws hold because

$$\begin{aligned} (f_1 \circ (f_2 + f_3))(r) &= f_1(f_2(r) + f_3(r)) \\ &= f_1(f_2(r)) + f_1(f_3(r)) \\ &= (f_1 \circ f_2)(r) + (f_1 \circ f_3)(r) \\ &= ((f_1 \circ f_2) + (f_1 \circ f_3))(r) \end{aligned}$$

for all $r \in R$. The right distributivity law follows by a very similar argument.

Finally, $\text{End}(R)$ has a multiplicative identity, the identity endomorphism $\text{id}: R \rightarrow R$.

More generally, we can construct endomorphism rings of other objects. The next example will resurface a little later when we discuss subrings.

Example 8.24 Let A be an abelian group. Then the set $\text{End}(A)$ of all abelian group endomorphisms $f: A \rightarrow A$ forms a ring under the addition and composition operations discussed earlier.

He cited the quadratic equation as an example of the sort of irrelevant topic that pupils study. I had hoped that the Government would make a robust rebuttal, but there was no defence either of mathematics in general or the quadratic equation in particular.

— Tony McWalter MP,
House of Commons debate on
Quadratic Equations. Hansard HC
Deb 26 June 2003 407:1259

8.3 Polynomials

SO FAR, we've met a number of examples of rings, commutative and noncommutative, unital and nonunital, formed from numbers and matrices of various types. The next simplest class of mathematical objects that most of us meet are polynomial expressions. We typically meet these in secondary school, initially as linear or quadratic expressions such as $3x - 2$ or $x^2 - 4x + 3$, and quickly learn how to add and multiply them. We also learn how to find their roots, but we'll leave that aspect for the moment and concentrate just on their additive and multiplicative behaviour.

Addition of polynomials is associative and commutative, there is a zero polynomial $0 = 0 + 0x + 0x^2 + \cdots$ which leaves unchanged any

other polynomial we add it to, and every polynomial $p(x)$ has a unique negative counterpart $-p(x)$. So polynomials certainly exhibit an abelian group structure.

We can also multiply polynomials together; this operation is associative and has an identity element $1 = 1 + 0x + 0x^2 + \cdots$. Furthermore, addition and multiplication of polynomials obeys the left and right distributive rules. This leads us to the following key example.

Example 8.25 Let x be a formal variable (an **unknown** or **indeterminate**). The set

$$\mathbb{Z}[x] = \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n : a_0, \dots, a_n \in \mathbb{Z}, n \in \mathbb{Z}_{\geq 0}\}$$

forms a commutative ring under the usual addition and multiplication operations.

There is an important detail that we must be careful not to gloss over at this point. Each polynomial in $\mathbb{Z}[x]$ is a $\mathbb{Z}$ -linear combination of nonnegative powers of the variable x , and must therefore have only a finite number of nonzero terms. To make this a little more precise, we introduce the following definition.

Definition 8.26 Let $p(x) = a_0 + a_1x + \cdots + a_nx^n$ be a polynomial in $\mathbb{Z}[x]$ such that $n \neq 0$. The **degree** of p , denoted $\deg(p)$, is the highest nonzero power of the variable x represented; that is, $\deg(p) = n$.

The degree of a quadratic polynomial $a_2x^2 + a_1x + a_0$ is therefore 2, that of a linear polynomial $a_1x + a_0$ is 1, and the degree of a constant element $a_0 \in \mathbb{Z}[x]$ is 0.

The degree of the zero polynomial 0 is usually left undefined, although some books set it by convention to either -1 or $-\infty$.

One generalisation of Example 8.25 is to let R be any ring and let

$$R[x] = \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n : a_0, \dots, a_n \in R, n \in \mathbb{Z}_{\geq 0}\}.$$

denote the ring of polynomials in x with coefficients in R . This ring $R[x]$ is commutative if and only if R is commutative, and unital if and only if R is unital. The rings $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ of polynomials with, respectively, integer and rational coefficients, are particularly interesting from an algebraic point of view: the search for general strategies for finding roots of polynomials in $\mathbb{Q}[x]$ led to the study of field extensions and the formulation of the machinery we now know as Galois Theory, named in memory of one of its principle discoverers. Polynomial rings of the form $\mathbb{Z}_p[x]$, where p is prime, have important applications in cryptography and coding theory.

We can allow negative powers of the variables as well; this leads to the following class of rings:

Example 8.27 Let $\mathbb{Z}[x^{\pm 1}]$ be the ring of polynomials of the form

$$a_{-m}x^{-m} + a_{-m+1}x^{-m+1} + \cdots + a_{-1}x^{-1} + a_0 + a_1x + \cdots + a_nx^n$$

with $a_{-m}, \dots, a_n \in \mathbb{Z}$. This is the ring of **Laurent polynomials** in x .

Laurent polynomials are of particular relevance in knot theory: the Alexander polynomial of a knot or link is an element of $\mathbb{Z}[t^{\pm 1}]$.

Rings of **multivariate polynomials** (polynomials with more than one unknown) play a fundamental rôle in the field of algebraic geometry. A detailed account is well beyond the scope of this book, but we can at least define them and study some of their basic properties.

Definition 8.28 Let R be a ring, let $x_1, \dots, x_k$ be unknowns, and let $n_1, \dots, n_k$ be nonnegative integers. Then a formal product of the form $x_1^{n_1} \cdots x_k^{n_k}$ is a **monomial**. A **(multivariate) polynomial** over R (or with **coefficients** in R) is an R -linear combination of monomials. The **degree** of a monomial is the sum $n_1 + \cdots + n_k$ of its exponents, and the degree of a given multivariate polynomial is the maximal degree of all its constituent monomials with nonzero coefficients.

We can now construct rings of multivariate polynomials.

Example 8.29 Let R be a ring, and let $x_1, \dots, x_k$ be formal variables. The set $R[x_1, \dots, x_k]$ of all R -linear combinations of finite-degree monomials in $x_1, \dots, x_k$ forms a ring under the usual addition and multiplication operations.

In particular, we assume that the unknowns $x_1, \dots, x_k$ commute with each other, so that

$$(ax_1^{n_1} \cdots x_k^{n_k}) \cdot (bx_1^{m_1} \cdots x_k^{m_k}) = (a \cdot b)x_1^{m_1+n_1} \cdots x_k^{m_k+n_k}.$$

Another way of generalising the ring $R[x]$ is to allow infinite-degree polynomials. This leads to the following example.

Example 8.30 Let x be an unknown and let R be a ring. Then a **power series** over R is a formal sum

$$\sum_{i=0}^{\infty} a_i x^i = a_0 + a_1 x + a_2 x^2 + \cdots$$

where $a_0, \dots \in R$. The set $R[[x]]$ of all power series over R forms a ring under the following addition and multiplication operations:

$$\begin{aligned} \left(\sum_{i=0}^{\infty} a_i x^i \right) + \left(\sum_{i=0}^{\infty} b_i x^i \right) &= \sum_{i=0}^{\infty} (a_i + b_i) x^i \\ \left(\sum_{i=0}^{\infty} a_i x^i \right) \cdot \left(\sum_{i=0}^{\infty} b_i x^i \right) &= \sum_{i=0}^{\infty} \left(\sum_{j=0}^i a_j \cdot b_{i-j} \right) x^i \end{aligned}$$

For polynomial rings we use single brackets, while for power series rings we use double brackets. So, for example, the ring of polynomials with integer coefficients is denoted $\mathbb{Z}[x]$, while the corresponding ring of power series is denoted $\mathbb{Z}[[x]]$.

You have probably met power series before, in the context of Taylor's

Theorem and Taylor–MacLaurin series, most likely in a course on calculus or analysis. When studying these objects from an analytical viewpoint, we must be careful to consider questions of convergence, and indeed there are a number of standard analytical results such as d’Alembert’s Ratio Test and the Alternating Series Test to answer these questions; now, however, we are viewing them purely as formal algebraic objects and can cheerfully sidestep such analytical concerns.

Nevertheless, the first time we meet polynomials and power series, we typically study them as functions: we are interested in what happens when we set the variable x to a particular value. We will return to this question in the next chapter, when we meet the **evaluation homomorphism** $f_a: R[x] \rightarrow R$ determined by a given element $a \in R$.³

³ Example 9.20, page 353.

We end this section with a class of rings that are in some sense a generalisation of the polynomial rings $R[x]$.

Example 8.31 Let G be a group, which for notational convenience we write multiplicatively. Then the **(integral) group ring** is the set

$$\mathbb{Z}G = \left\{ \sum_{g \in G} n_g g : \text{only finitely many } n_g \neq 0 \right\}$$

consisting of finite $\mathbb{Z}$ –linear combinations of elements of G .

The addition operation in $\mathbb{Z}G$ is the obvious one, namely

$$\sum_{g \in G} m_g g + \sum_{g \in G} n_g g = \sum_{g \in G} (m_g + n_g) g$$

and the multiplication operation is

$$\left(\sum_{g \in G} m_g g \right) \cdot \left(\sum_{g \in G} n_g g \right) = \sum_{g \in G} \left(\sum_{h * k = g} m_h + n_k \right) g.$$

That is, we multiply two sums in the usual way, combining terms by using the ordinary multiplication operation in $\mathbb{Z}$ and the group multiplication operation accordingly: $n_g g \cdot m_h h = (n_g \cdot m_h)(g * h)$.

We can generalise this by replacing the coefficient ring $\mathbb{Z}$ with some other ring R to get RG .

The group ring is important in representation theory and homological algebra. We will revisit it in Section 8.A, where we will see that modules over $\mathbb{Z}G$ are exactly the G –modules discussed in Section 7.5.⁴

⁴ Definition 7.90, page 279.

Example 8.32 Let $G = \langle t : t^2 = 1 \rangle \cong \mathbb{Z}_2$. Then $\mathbb{R}G$ consists of elements of the form $a + bt$ where $a, b \in \mathbb{R}$ and

$$\begin{aligned} (a + bt) + (c + dt) &= (a + c) + (b + d)t, \\ (a + bt) \cdot (c + dt) &= (ac + bd) + (ad + bc)t. \end{aligned}$$

This is the ring of **split complex**, **perplex** or **hyperbolic** numbers.

Division findeth how oft one Number is contained in another. The Number divided, (or which is consider'd as the containing number) is called the Dividend; the Number dividing (or which is considered as contained) is called the Divisor; and the Number sought, is called the Quotient, or Quote, (from Quoties, How oft;) because it shews the how oft sought.

— Alexander Malcolm (1685–1763),
*A New System of Arithmetick, Theoretical
 and Practical* (1730) 52

8.4 Fields

AT THE BEGINNING OF THIS CHAPTER we compiled a list of properties a multiplication operation might satisfy. So far, we've looked at the existence of an identity element, and also considered examples where the commutativity condition might or might not be satisfied. The other obvious question to ask is whether some or all elements of a ring have a multiplicative inverse.

For any nontrivial ring R , at least one element isn't invertible: the additive identity 0 , since there is no element $r \in R$ such that $r \cdot 0 = 1$, as $r \cdot 0 = 0$ for all $r \in R$ by Proposition 8.8 (i).

In the ring $\mathbb{Z}_2[i]$, the element $1+i$ doesn't have an inverse, but 1 and i both do. In $\mathbb{Z}_2[\omega]$, however, every element apart from 0 is invertible:

$$1^{-1} = 1, \quad \omega^{-1} = \omega^2, \quad (\omega^2)^{-1} = \omega.$$

So, in general, 0 isn't invertible (except in the case of the trivial ring), 1 is invertible (it's its own inverse) and some or all of the other elements might be invertible. We give invertible elements a special name:

Definition 8.33 Let R be a (unital) ring. If $r \in R$ is invertible, in the sense that there exists some element $s = r^{-1} \in R$ such that $r \cdot s = s \cdot r = 1$, then we say that r is a **unit** (and so is s).

The units of a ring have some interesting properties: they are closed under multiplication and include the identity, leading to the following.

Proposition 8.34 Let $U(R)$ be the set of units of a (unital) ring R . Then $U(R)$ forms a group, the **group of units**, under multiplication.

Proof The unity element $1 \in R$ is contained in $U(R)$ since $1^{-1} = 1$. For any two units $a, b \in U(R)$ the product $a \cdot b$ is also in $U(R)$ since $a \cdot b$ has inverse $(a \cdot b)^{-1} = b^{-1} \cdot a^{-1}$. If $a \in U(R)$ then its inverse a^{-1} must also be in $U(R)$. Finally, multiplication in $U(R)$ is associative because it is in R . Hence $U(R)$ is a group. $\square$

We've met this concept already, back in Section 2.3,^{5,6} when we studied the multiplicative groups $\mathbb{Z}_n^\times$, which are exactly the groups $U(\mathbb{Z}_n)$.⁷

Example 8.35 The ring $\mathbb{Z}$ of integers has only two invertible elements: 1 and -1 . Hence $U(\mathbb{Z}) = \{\pm 1\} \cong \mathbb{Z}_2$.

Example 8.36 The polynomial ring $\mathbb{Z}[t]$ also has only two invertible elements: 1 and -1 . Thus $U(\mathbb{Z}[t]) = \{\pm 1\} \cong \mathbb{Z}_2$

The ring $\mathbb{Z}[t^{\pm 1}]$ of Laurent polynomials over $\mathbb{Z}$ has infinitely many units: any element of the form $\pm t^n$, where $n \in \mathbb{Z}$, is invertible (its inverse is $\mp t^{-n}$). Hence $U(\mathbb{Z}[t^{\pm 1}]) = \{\pm t^n : n \in \mathbb{Z}\} \cong \mathbb{Z}_2 \oplus \mathbb{Z}$.

⁵ Definition 2.41, page 62.

⁶ Example 2.42, page 62.

⁷ Some books denote $U(R)$ by $R^\times$ for this reason.

Example 8.37 The invertible elements in the ring $M_n(\mathbb{R})$ are exactly the $n \times n$ real matrices with nonzero determinant. Hence $U(M_n(\mathbb{R})) = GL_n(\mathbb{R})$.

Example 8.38 The ring $\mathbb{Q}$ of rational numbers consists of all quotients of the form $\frac{a}{b}$ where $b \neq 0$. The multiplicative inverse of $\frac{a}{b}$ is obviously $\frac{b}{a}$, which is only defined if $a \neq 0$. Hence $U(\mathbb{Q}) = \mathbb{Q}^*$, the multiplicative group of nonzero rational numbers.

Similarly, $U(\mathbb{R}) = \mathbb{R}^*$ and $U(\mathbb{C}) = \mathbb{C}^*$.

This last example displays the best case scenario: everything is a unit except for 0 (which definitely can't be). Rings satisfying this criterion are particularly versatile to work with: we can divide arbitrary nonzero elements by each other as well as adding, subtracting and multiplying them. We can legitimately ask (and answer) a wider range of questions in a ring with this property, such as "what gives 1 when multiplied by 3?". It therefore makes sense to give this concept a specific name.

Definition 8.39 A **division ring** is a ring R such that $U(R) = R^* = R \setminus \{0\}$. That is, every nonzero element $r \in R$ is a unit.

This isn't to say that rings without this property aren't interesting. On the contrary, such rings often display interestingly awkward behaviour. Even nicer is the case where such a ring is commutative as well:

Definition 8.40 A **field** is a commutative division ring. A noncommutative division ring is called a **skew field**.

Many of the number systems we're most familiar with are fields. In fact, in some sense the point of the rational numbers $\mathbb{Q}$ is that they are the integers $\mathbb{Z}$ made as invertible as possible. We'll come back to this idea in a little while when we meet the concept of a **field of quotients**, but for the moment here are some examples of fields and skew fields.

Example 8.41 The rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$ are all fields.

The quaternions $\mathbb{H}$ form a skew field.

We met another important class of fields in Chapter 3:⁸

⁸ Definition 3.45, page 94.

Example 8.42 Let p be prime. Then the set $\mathbb{F}_p = \{0, \dots, p-1\}$ forms a field (the **finite field** of order p) under addition and multiplication modulo p .

In a little while we will prove that $\mathbb{F}_p$ is indeed a field. Finite fields have particularly important applications in cryptography and coding theory. Some forms of public key encryption rely on the computational difficulty of factorising composite numbers over finite fields, while

many useful error-correcting codes are defined in terms of vector subspaces over finite fields. Later, in Chapter 11, we will see that finite fields of order $q = p^n$ both exist and are unique up to isomorphism, for any prime p and $n \in \mathbb{N}$.

After learning basic arithmetic, one of the next things we learn how to do is to factorise simple polynomials and thereby find their roots.

For example, the quadratic polynomial $p(x) = x^2 - 3x + 2$ can be factorised as $(x - 1)(x - 2)$, at which point we can readily see that its roots (the values of x for which $p(x) = 0$) are 1 and 2. In doing so we are making use of a fundamental property of the number system we're working in, whether that be $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$.

Usually, for $(x - 1)(x - 2)$ to equal zero it must follow that either $(x - 1) = 0$, in which case $x = 1$, or $(x - 2) = 0$, in which case $x = 2$. But we have already met several examples of rings in which this doesn't hold. For example, in the ring $\mathbb{Z}_6$ (see Table 8.6) we have $2 \cdot 3 = 0$. So although it's always true (from Proposition 8.8(i)) that $a \cdot b = 0$ if either $a = 0$ or $b = 0$, there are certainly some rings in which the converse doesn't hold.

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

·	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Table 8.6: Addition and multiplication tables for the ring $\mathbb{Z}_6$

Definition 8.43 Let $a \in R$ be a nonzero element of some ring R . Then a is a **left zero divisor** if there exists some $b \in R$ such that $a \cdot b = 0$, and a **right zero divisor** if there exists some $c \in R$ such that $c \cdot a = 0$.

If a is both a left and right zero divisor then it is said to be a **two-sided zero divisor**, or simply a **zero divisor**.

Using this terminology, we see from Table 8.6 that $2, 3, 4 \in \mathbb{Z}_6$ are all zero divisors, but 1 and 5 aren't (and by convention neither is 0).

More generally, there is a simple criterion governing whether a particular nonzero element of a ring $\mathbb{Z}_n$ is a zero divisor:

Proposition 8.44 An element $m \in \mathbb{Z}_n$ is a zero divisor exactly when it is not coprime to n ; that is, when $\gcd(m, n) > 1$.

Proof Suppose that $\gcd(m, n) = d > 1$. Then

$$m \cdot \frac{n}{d} = \frac{m}{d} \cdot n \equiv 0 \pmod{n}.$$

Hence $m \cdot \frac{n}{d} = 0$ in $\mathbb{Z}_n$, but neither $m = 0$ nor $\frac{n}{d} = 0$, so they are both zero divisors.

Conversely, suppose that $\gcd(m, n) = 1$, and that there exists some $k \in \mathbb{Z}_n$ for which $mk = 0$. This means that n divides mk , and hence by Proposition A.40, since m and n are coprime, n must divide k instead. Thus $k \equiv 0 \pmod{n}$, which means that $k = 0$ in $\mathbb{Z}_n$, and therefore m is not a zero divisor. $\square$

Matrix rings, the source of so many interesting examples, also yield

instances of zero divisors:

Example 8.45 Let $A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \in M_2(\mathbb{Z})$. Then A is a left zero divisor via $B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ and a right zero divisor via $C = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

The property that $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ have that simplifies the process of finding roots of polynomials, is that there are no nonzero elements that multiply to give zero. That is, they have no zero divisors.

Definition 8.46 An **integral domain** is a commutative (unital) ring with no zero divisors.

Clearly not every integral domain is a field, since $\mathbb{Z}$ has the required property too, but it so happens that every field is an integral domain:

Proposition 8.47 Let F be a field. Then F has no zero divisors, and is therefore an integral domain.

Proof Suppose that $a, b \in F$ such that $a \cdot b = 0$ and $a \neq 0$. Then

$$b = 1 \cdot b = (a^{-1} \cdot a) \cdot b = a^{-1} \cdot (a \cdot b) = a^{-1} \cdot 0 = 0.$$

Hence F has no zero divisors. □

An important aspect of integral domains is that although they needn't contain a full complement of multiplicative inverses, they do satisfy left and right cancellation laws analogous to those in Proposition 1.15.

Proposition 8.48 Let R be an integral domain. Then for any $a, b, c \in R$ such that $a \neq 0$ and $a \cdot b = a \cdot c$ it follows that $b = c$.

Similarly, for any nonzero $d \in R$ such that $b \cdot d = c \cdot d$ it follows that $b = c$.

Proof Suppose that $a \cdot b = a \cdot c$. Then $a \cdot b - a \cdot c = a \cdot (b - c) = 0$, and since R , as an integral domain, has no zero divisors, and since $a \neq 0$, it follows that $(b - c) = 0$ and hence $b = c$. The right-hand cancellation law follows by a similar argument. □

Proposition 8.47 tells us that all fields are integral domains, but obviously not all integral domains are fields: for example, neither $\mathbb{Z}$ nor the polynomial rings $\mathbb{Z}[t]$ or $\mathbb{Z}[t^{-1}]$ are fields, but are certainly integral domains. These examples are infinite, though, so can we find any finite integral domains that aren't fields? The answer is no:

Proposition 8.49 Let R be a finite integral domain. Then R is a field.

Proof Let $R = \{0, r_1, \dots, r_n\}$ be a finite integral domain. Choose $1 \leq i \leq n$ and consider the n products $r_i \cdot r_1, \dots, r_i \cdot r_n$. By the left cancellation law in Proposition 8.48 these must all be distinct, and since R is an integral domain none of them can be equal to zero.

Exactly one of these products $r_i \cdot r_j$ must be equal to 1, and hence $r_j = r_i^{-1}$. Repeating this argument $n-1$ further times, we find that

every $r_i \in R$ has a unique inverse, and hence R is a field. $\square$

Corollary 8.50 *If p is prime, then $\mathbb{Z}_p$ is a field.*

Proof By Proposition 8.44, the zero divisors in $\mathbb{Z}_n$ are those integers $k \in \mathbb{Z}_n$ which are not coprime to n ; that is, for which $\gcd(k, n) > 1$. Since p is prime, there are no such integers $k \in \mathbb{Z}_p$ and hence $\mathbb{Z}_p$ is an integral domain. Furthermore, it has finitely many elements, and by Proposition 8.49 must be a field. $\square$

The first field most of us encounter is $\mathbb{Q}$, the field of rational numbers, although we don't usually know it by that name to start with. Typically, after learning about addition, subtraction and multiplication of integers, we learn about division. In many cases, one integer will divide exactly by another, leaving no remainder, but often it won't. In order to deal with such an eventuality, at least in circumstances where this might be important to us, we extend our concept of a number by assigning meaning to expressions such as $\frac{1}{7}$ and $\frac{355}{113}$. In effect, we construct a new ring containing $\mathbb{Z}$ as a unital subring by allowing expressions of this type; in particular we get a multiplicative inverse $n^{-1} = \frac{1}{n}$ for any nonzero integer $n \in \mathbb{Z}$.

We now want to generalise this construction to work with any integral domain, and to do so we will look carefully at exactly what's going on with the construction of the rationals from the integers.

To start with, every rational number $q \in \mathbb{Q}$ can be expressed as a fraction $\frac{a}{b}$ where $a, b \in \mathbb{Z}$ and $b \neq 0$. Also, we want a and b to be coprime, so that $\frac{a}{b}$ is in its simplest form. This is the same as requiring $\gcd(a, b) = 1$, which is fine for integers, but we want a construction that works in any integral domain, and we haven't yet decided what, if anything, we mean by a greatest common divisor in general.⁹

⁹In the next chapter we'll return to this question, in the process discovering that although we can certainly define a more general notion of greatest common divisor, they don't necessarily exist in an arbitrary integral domain.

We want to ensure that every element of our new field is well-defined; for example, we regard $\frac{1}{2}$ to be the same as $\frac{2}{4}$. Given that we can't rely on greatest common divisors or coprimality, we need a slightly different approach. The answer is to define a suitable equivalence relation on this larger set and then mod out by that relation.

Suppose that $\frac{a}{b} = \frac{c}{d}$. Then by cross-multiplying, we see that this is equivalent to saying that $ad = bc$. So, we form the field $\mathbb{Q}$ of rational numbers by taking the Cartesian product

$$\mathbb{Z} \times \mathbb{Z}_* = \{(a, b) : a, b \in \mathbb{Z}, b \neq 0\}$$

and then quotienting by the relation $\sim$ where $(a, b) \sim (c, d)$ if $ad = bc$.

The equivalence classes determined by this relation are called **fractions**, and we usually write them as $\frac{a}{b}$ or a/b .

The next thing we need to do is to decide how to add and multiply

two fractions together, and also figure out how the ring $\mathbb{Z}$ embeds in this field.

Addition and multiplication is given by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd},$$

and the integers $\mathbb{Z}$ embed via the homomorphism $i: \mathbb{Z} \hookrightarrow \mathbb{Q}$ given by $f(n) = \frac{n}{1}$.

All of this works fine in any integral domain, leading to the following.

Definition 8.51 Let R be an integral domain, and let

$$Q(R) = (R \times R_{\neq 0}) / \sim$$

be the set of equivalence classes of the relation $\sim$ defined by $(a, b) \sim (c, d)$ when $ad = bc$. We represent the equivalence class $[(a, b)]_{\sim}$ by a/b or $\frac{a}{b}$. Then $Q(R)$ forms a field (the **field of fractions** or **field of quotients**) with the addition and multiplication operations

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

for any $a, b, c, d \in R$ such that $b, d \neq 0$.

Furthermore, there is an injective ring homomorphism $i: R \rightarrow Q(R)$ mapping $r \mapsto r/1$ for all $r \in R$.

So, the field of fractions $Q(\mathbb{Z})$ of the integers $\mathbb{Z}$ is the field $\mathbb{Q}$ of rational numbers. This shouldn't be surprising, because we formulated the definition of $Q(R)$ precisely so that this would be the case.

Example 8.52 Let's calculate the field of fractions $Q(\mathbb{Z}_3)$ of the ring $\mathbb{Z}_3$. First, we take the set

$$\mathbb{Z}_3 \times (\mathbb{Z}_3 \setminus \{0\}) = \{(0, 1), (1, 1), (2, 1), (0, 2), (1, 2), (2, 2)\}.$$

Now we mod out by the equivalence relation $\sim$ to find that

$$(0, 1) \sim (0, 2), \quad (1, 1) \sim (2, 2), \quad (1, 2) \sim (2, 1).$$

Hence

$$Q(\mathbb{Z}_3) = \left\{ \frac{0}{1}, \frac{1}{1}, \frac{2}{1} \right\}$$

which is just the image of $\mathbb{Z}_3$ under the inclusion homomorphism $i: \mathbb{Z}_3 \hookrightarrow Q(\mathbb{Z}_3)$. Hence $Q(\mathbb{Z}_3) \cong \mathbb{Z}_3$.

This makes sense, since $\mathbb{Z}_3$ is already a field under the usual modulo-3 addition and multiplication operations. More generally:

Proposition 8.53 Let $\mathbb{F}$ be a field. Then $Q(\mathbb{F}) \cong \mathbb{F}$.

Proof The required isomorphism is the inclusion homomorphism $i: \mathbb{F} \hookrightarrow Q(\mathbb{F})$. We already know this is injective, so we just need to show it's surjective as well. Any fraction $\frac{a}{b} \in Q(\mathbb{F})$ can be expressed

as the image $i(ab^{-1})$ of the element $ab^{-1} \in \mathbb{F}$, since

$$i(ab^{-1}) = i(a) \cdot i(b^{-1}) = \frac{a}{1} \cdot \frac{b^{-1}}{1} = \frac{a}{1} \cdot \frac{1}{b} = \frac{a}{b}.$$

Hence i is surjective as well, and therefore an isomorphism. $\square$

Example 8.54 What is the field of fractions $Q(\mathbb{Z}[i])$ of the ring $\mathbb{Z}[i]$ of Gaussian integers? It consists of all fractions of the form

$$\frac{a + bi}{c + di}$$

where $a, b, c, d \in \mathbb{Z}$, and $c + di \neq 0$.

Using the usual trick for complex division, we find that

$$\frac{a + bi}{c + di} = \frac{a + bi}{c + di} \cdot \frac{c - di}{c - di} = \frac{a}{c^2 + d^2} + \frac{b}{c^2 + d^2}i.$$

Hence, $Q(\mathbb{Z}[i]) = \{p + qi : p, q \in \mathbb{Q}\} = \mathbb{Q}[i]$.

The following example is one that we'll make use of in Chapter 11.

Example 8.55 The field of fractions $Q(\mathbb{Z}[x])$ of the polynomial ring $\mathbb{Z}[x]$ consists of expressions of the form

$$\frac{a_mx^m + \cdots + a_0}{b_nx^n + \cdots + b_0}$$

where $a_0, \dots, a_m, b_0, \dots, b_n \in \mathbb{Z}$ and $b_nx^n + \cdots + b_0 \neq 0$. These are called **rational functions**, and we denote the resulting field $Q(x)$. (Note the use of parentheses instead of brackets.)

The main point of the field of fractions is that it's the smallest field containing the given integral domain.

8.A Modules and representations

Summary

We began this chapter with another careful examination of the integers, this time paying attention not just to the additive structure, but the multiplicative structure as well. This led us to formulate the concept of a **ring**: a set $R = (R, +, \cdot)$ equipped with two binary operations $+$ and $\cdot$.¹⁰ We require that $(R, +)$ is an abelian group, and that the multiplication operation $\cdot$ is associative and satisfies a distributivity condition with respect to addition. We also require the existence of a unique¹¹ **unity** element which acts as a multiplicative identity; a ring without such an element is called a **ring without unity** or a

¹⁰ Definition 8.1, page 318.

¹¹ Proposition 8.9, page 321.

nonunital ring. (To avoid ambiguity, we sometimes call a ring with a multiplicative identity a **unital ring** or a **ring with unity**, but in this book we will consider this to be the default setting.)

The first examples we looked at were numerical ones: the integer rings $\mathbb{Z}$ and $\mathbb{Z}_n$, the rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$ and the complex numbers $\mathbb{C}$.^{12,13} The simplest ring is the **trivial ring** consisting of a single element, which must be both the additive and multiplicative identity;¹⁴ this is the only ring in which $0 = 1$.¹⁵ We can also make new rings by adjoining a new element to an existing one; this led us to the ring $\mathbb{Z}[i]$ of **Gaussian integers**, the ring $\mathbb{Z}[\omega]$ of **Eisenstein integers** and the ring $\mathbb{Z}[\lambda]$ of **Kleinian integers**,¹⁶ as well as some finite variants.¹⁷ Several basic properties may be quickly deduced from the definition of a ring; in particular multiplication by the zero (additive identity) element and negative (additive inverse) elements behaves in the way we're used to from ordinary arithmetic.¹⁸

Although we'll mostly be interested in unital rings, we also looked at a few examples of nonunital rings, such as the even integers $2\mathbb{Z}$ (which generalises to $n\mathbb{Z}$ in the obvious way)¹⁹ and the case where we define multiplication to be trivial.²⁰ The **quaternions** $\mathbb{H}$ and the **Lipschitz integers** $\mathbb{L}$ provided interesting examples of noncommutative rings.²¹

In order to better study structural similarities between rings, we introduced the notions of an **isomorphism** and a **homomorphism**,²² and **direct sum**, in both cases extending the corresponding concepts from group theory. As with group homomorphisms, any ring homomorphism must map 0 to 0 and 1 to 1.²³ Next we considered the **additive** and **multiplicative order** of ring elements,²⁴ and introduced the **characteristic** $\text{char}(R)$ of a ring,²⁵ which latter may be determined from the additive order of the unity element 1.²⁶

Next we examined some examples of matrix rings such as $M_n(R)$, the ring of $n \times n$ matrices with entries from some ring R ,²⁷ as well as the rings $UT_n(R)$, $LT_n(R)$ and $D_n(R)$ of triangular and diagonal matrices.²⁸ The ring $M_n(R)$ is noncommutative except when $n = 1$ and R is commutative, and its elements can be regarded as endomorphisms of R ; more generally we can form the ring $\text{End}(R)$ of endomorphisms of R ,²⁹ and also generalise this to endomorphism rings of other objects such as abelian groups.³⁰

The next class of rings we studied were polynomial rings such as $\mathbb{Z}[x]$.³¹ These are formed by adjoining a formal variable x to an existing ring, in a similar way to the construction of rings such as $\mathbb{Z}[i]$ and $\mathbb{Z}_2[\sqrt{2}]$. This process can be generalised in various ways: we can use an arbitrary coefficient ring to get $R[x]$, allowing negative powers yields the ring $\mathbb{Z}[x^{\pm 1}]$ of **Laurent polynomials**,³² we can define mul-

¹² Example 8.2, page 319.

¹³ Example 8.3, page 319.

¹⁴ Example 8.4, page 319.

¹⁵ Proposition 8.13, page 323.

¹⁶ Example 8.6, page 320.

¹⁷ Example 8.7, page 320.

¹⁸ Proposition 8.8, page 320.

¹⁹ Example 8.10, page 322.

²⁰ Example 8.11, page 322.

²¹ Example 8.12, page 322.

²² Definition 8.14, page 323.

²³ Proposition 8.16, page 324.

²⁴ Definition 8.17, page 325.

²⁵ Definition 8.18, page 325.

²⁶ Proposition 8.19, page 325.

²⁷ Example 8.21, page 326.

²⁸ Example 8.22, page 327.

²⁹ Example 8.23, page 328.

³⁰ Example 8.24, page 328.

³¹ Example 8.25, page 329.

³² Example 8.27, page 330.

³³ Example 8.29, page 330.³⁴ Example 8.30, page 330.³⁵ Example 8.31, page 331.³⁶ Definition 8.33, page 332.³⁷ Proposition 8.34, page 332.³⁸ Definition 8.39, page 333.³⁹ Definition 8.40, page 333.⁴⁰ Corollary 8.50, page 336.⁴¹ Definition 8.43, page 334.⁴² Proposition 8.44, page 334.⁴³ Definition 8.46, page 335.⁴⁴ Proposition 8.47, page 335.⁴⁵ Proposition 8.49, page 335.⁴⁶ Proposition 8.48, page 335.

tivariate polynomial rings like $\mathbb{Z}[x, y]$ or $R[x_1, \dots, x_k]$ by using more than one variable,³³ and by allowing infinitely many terms we obtain **formal power series rings** like $\mathbb{Z}[[x]]$.³⁴ A related example is that of the **group ring** $\mathbb{Z}G$ (or RG) of a group G , whose elements are formal $\mathbb{Z}$ -linear (or R -linear) combinations of elements of G .³⁵

Next we considered the multiplicative structure of rings in more detail, introducing the notion of an invertible element, or **unit**.³⁶ The units in a ring R are closed under multiplication, include the unity element 1 itself, and all have unique inverses. They therefore form a multiplicative group, the **group of units** $U(R)$.³⁷ The rings $\mathbb{Z}$ and $\mathbb{Z}[t]$ have only two units, namely ± 1 , and hence $U(\mathbb{Z}) = U(\mathbb{Z}[t]) \cong \mathbb{Z}_2$. The ring $\mathbb{Z}[t^{\pm 1}]$ of Laurent polynomials has infinitely many units of the form $\pm t^n$ for $n \in \mathbb{Z}$ and hence $U(\mathbb{Z}[t^{\pm 1}]) \cong \mathbb{Z}_2 \oplus \mathbb{Z}$. The invertible elements in $M_n(\mathbb{R})$ are exactly the $n \times n$ invertible matrices, and thus $U(M_n(\mathbb{R})) = GL_n(\mathbb{R})$.

The best case scenario is that every nonzero element in a ring R is invertible (the zero element 0 is never invertible unless $R = \{0\}$ is the trivial ring). We call such a ring a **division ring**.³⁸ A commutative division ring is a **field** and a noncommutative division ring is a **skew field**.³⁹ Many familiar number systems are fields: the rationals $\mathbb{Q}$, the reals $\mathbb{R}$ and the complex numbers $\mathbb{C}$. The quaternions $\mathbb{H}$ are perhaps the best-known example of a skew field. For a prime p , the set $\mathbb{F}_p = \{0, 1, \dots, p-1\}$ forms a **finite field** under modulo- p addition and multiplication.⁴⁰

A nonzero element a in a ring R is a **left zero divisor** if there exists some nonzero $b \in R$ with $a \cdot b = 0$, and a **right zero divisor** if there exists some nonzero $c \in R$ such that $c \cdot a = 0$. If a is both a left and right zero divisor then we call it simply a **zero divisor**.⁴¹ The zero divisors in $\mathbb{Z}_n$ are exactly the elements m that aren't coprime to n ; that is, those for which $\gcd(m, n) > 1$.⁴² A ring such as $\mathbb{Z}$ or $\mathbb{Z}[t]$ that contains no zero divisors is said to be an **integral domain**.⁴³ All fields are integral domains,⁴⁴ and all *finite* integral domains are fields,⁴⁵ but not all *infinite* integral domains are fields. Although integral domains don't necessarily contain a full complement of units, they do satisfy multiplicative cancellation laws.⁴⁶

References and further reading

The mathematical biographer Constance Reid (1918–2010) wrote an interesting and well-researched biography of David Hilbert (1862–1943) which contains many details of his life and also of the mathematical community at Göttingen during the late 19th and early 20th century.

- C Reid, *Hilbert*, Copernicus (1996)

The German geologist Wolfgang Sartorius von Waltershausen (1809–1876) wrote a contemporary biography of Carl Friedrich Gauss (1777–1855):

- WS von Waltershausen, *Gauss zum Gedächtnis*, S Hirzel (1856)

An English translation by Helen W Gauss, the subject's great granddaughter, was published in 1966 as *Gauss: A Memorial*.

For further details on representation theory, the following two books are good places to start:

- G James and M Liebeck, *Representations and Characters of Groups*, second edition, Cambridge University Press (2001)
- W Fulton and J Harris, *Representation Theory: A First Course*, Graduate Texts in Mathematics 129, Springer (2004)

The former is suitable for undergraduate students, and covers group representations and characters, while the latter is intended for graduate students, and covers the representation theory of finite groups, and also Lie groups and Lie algebras.

Exercises

- 8.1 Show that the prime subring $\langle 1 \rangle$ of a ring R is equal to the intersection of all other subrings of R .
- 8.2 Write down the addition and multiplication tables for the rings $\mathbb{Z}_2[\lambda]$, $\mathbb{Z}_3[i]$, $\mathbb{Z}_3[\sqrt{2}]$, $\mathbb{Z}_3[\omega]$ and $\mathbb{Z}_3[\lambda]$.
- 8.3 Show that if $f: R \rightarrow S$ is a homomorphism of unital rings, then there is a unique induced homomorphism $U(f): U(R) \rightarrow U(S)$, between the corresponding groups of units, such that $U(f)(r) = f(r)$ for all $r \in U(R)$.
- 8.4 A ring element $r \in R$ is **idempotent** if $r^2 = r$. Show that if R is an integral domain, then 0 and 1 are the only idempotent elements.
- 8.5 Let R be an integral domain. Show that the field of fractions $Q(R)$ from Definition 8.51 is indeed a field.
- 8.6 Let

$$H = \{a + bi + cj + dk : a, b, c, d \in \mathbb{Z} \text{ or } a, b, c, d \in \mathbb{Z} + \frac{1}{2}\}$$

be the set of **Hurwitz integers** or **Hurwitz quaternions**, where

$$\mathbb{Z} + \frac{1}{2} = \left\{ \dots, -\frac{3}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \dots \right\}$$

is the set of **half-integers**. Show that H is a noncommutative ring.

- 8.7 Let $\text{End}(\mathbb{Z})$ and $\text{End}(\mathbb{Z}_n)$ denote the endomorphism rings of the abelian groups $\mathbb{Z}$ and $\mathbb{Z}_n$. Show

that $\text{End}(\mathbb{Z}) \cong \mathbb{Z}$ and $\text{End}(\mathbb{Z}_n) \cong \mathbb{Z}_n$.

- 8.8 Show that $\mathbb{Z}[x][y] \cong \mathbb{Z}[x, y]$. That is, show that the ring of polynomials in y with coefficients in $\mathbb{Z}[x]$ is isomorphic to the ring of bivariate polynomials in x and y with coefficients in $\mathbb{Z}$.
- 8.9 Show that the integral group ring $\mathbb{Z}\mathbb{Z}$ is isomorphic to the Laurent polynomial ring $\mathbb{Z}[x^{\pm 1}]$.
- 8.10 Suppose that $\mathbb{Z}[i]$ is the ring of Gaussian integers from Example 8.6. Show that $U(\mathbb{Z}[i]) \cong \mathbb{Z}_4$.
- 8.11 Determine the group of units $U(M_n(\mathbb{Z}))$ of the ring of $n \times n$ integer matrices.

9 Ideals

AT ROUGHLY THIS POINT in our study of groups we met the concepts of subgroups and their left and right cosets, and fairly soon after we derived the notion of a **normal subgroup**: a subgroup whose left and right cosets coincide. We found that we could use these normal subgroups to form quotient groups, a process which gave us new structural information about the groups in question. We discovered a little while later that kernels of group homomorphisms are normal, and conversely that all normal subgroups arise in this way.

Now we want to study the analogous concepts in ring theory: subrings, cosets, kernels, homomorphisms, normal subrings and quotient rings. This will keep us busy for the rest of this chapter.

In science, nothing that is provable should be accepted without proof. As obvious as this requirement seems, yet I do not think it has been satisfied, even by the most recent methods in the foundations of the simplest science, namely that part of logic which concerns the theory of numbers.

— Richard Dedekind (1831–1916),
Was sind und was sollen die Zahlen?
(1888)

+	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,1)	(1,0)	(1,1)
(0,1)	(0,1)	(0,0)	(1,1)	(1,0)
(1,0)	(1,0)	(1,1)	(0,0)	(0,1)
(1,1)	(1,1)	(1,0)	(0,1)	(0,0)

·	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	(0,0)	(0,0)	(0,0)	(0,0)
(0,1)	(0,0)	(0,1)	(0,0)	(0,1)
(1,0)	(0,0)	(0,0)	(1,0)	(1,0)
(1,1)	(0,0)	(0,0)	(1,0)	(1,1)

Table 9.1: Addition and multiplication tables for the ring $\mathbb{Z}_2 \oplus \mathbb{Z}_2$

9.1 Subrings

IN THE EARLIER PART OF THIS BOOK, we learned a lot about the internal structure of groups by studying their subgroups: subsets that form groups in their own right with the same binary operation (or, more precisely, the restriction of the same binary operation).

We want now to formulate the corresponding concept for rings: a subset S of a ring R that forms a ring in its own right with the (restrictions of) the addition and multiplication operations inherited from R .

Looking at the addition and multiplication tables for $\mathbb{Z}_2 \oplus \mathbb{Z}_2$, in Table 9.1, we can see that this ring contains a subset $\{(0,0), (1,1)\}$ that forms a ring isomorphic to $\mathbb{Z}_2$. Furthermore, this subset contains both the additive and multiplicative identities.

The ring $\mathbb{Z}_2[i] \cong \mathbb{Z}_2[\sqrt{2}]$ also contains a subring isomorphic to $\mathbb{Z}_2$, as can be seen from Table 9.2.

Turning our attention to infinite rings, we can see that the ring $\mathbb{Q}$ of rational numbers contains $\mathbb{Z}$ as a ring in its own right, and similarly

Thus I descended out of the first circle
Down to the second, that less space
begirds,
And so much greater dole, that goads
to wailing.

— Dante Alighieri (c.1265–1321),
Inferno V:1–3,
translated by Henry Wadsworth
Longfellow (1807–1882)

+	0	1	1+i	i
0	0	1	1+i	i
1	1	0	i	1+i
1+i	1+i	i	0	1
i	i	1+i	1	0

·	0	1	1+i	i
0	0	0	0	0
1	0	1	1+i	i
1+i	0	1+i	0	1+i
i	0	i	1+i	1

Table 9.2: Addition and multiplication tables for the ring $\mathbb{Z}_2[i]$

$\mathbb{R}$ contains $\mathbb{Q}$, $\mathbb{C}$ contains $\mathbb{R}$, $\mathbb{Q}$ and $\mathbb{Z}$, and $\mathbb{H}$ contains all of these in turn.

What about the subset $2\mathbb{Z}$ of even integers in $\mathbb{Z}$? This is certainly closed under addition and contains both the additive identity 0 and all the positive and negative even integers, so it certainly forms an abelian group under addition. Multiplication of even integers is associative (and commutative) and satisfies the distributive law. But we're missing the remaining crucial ingredient: the multiplicative identity. So, as discussed in Example 8.10, it only forms a nonunital ring. We need to take account of this possibility in our definition of a subring.

At this point we run into some notational awkwardness. Just as different books adopt different conventions regarding whether rings have a unity element or not, there is no universally-observed convention concerning the definition of a subring.

Books that require rings to have a unity element by default usually also require subrings to have a unity element unless otherwise specified, while those that don't necessarily expect rings to have a unity element naturally tend not to expect subrings to either. Since we've decided to follow the "with unity" convention, our way should be clear. Unfortunately, however, things turn out to be more complicated, as the following example shows.

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

·	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Table 9.3: Addition and multiplication tables for $\mathbb{Z}_6$

Example 9.1 The ring $\mathbb{Z}_6$ has the multiplication and addition tables shown in Table 9.3. We can see from these that $\mathbb{Z}_6$ has a unity element, namely 1. Now consider the subset $\{0, 2, 4\}$. This forms an abelian group under modulo-6 addition and is closed under modulo-6 multiplication. Its addition and multiplication tables are

+	0	2	4
0	0	2	4
2	2	4	0
4	4	0	2

·	0	2	4
0	0	0	0
2	0	4	2
4	0	2	4

and from the second of these we can see that 4 acts as a unity element. Similarly, the subset $\{0, 3\}$ forms an abelian subgroup under addition, is closed under multiplication, and the element 3 acts as a multiplicative identity.

In neither of these cases is the unity element in the subset the same as the unity element $1 \in \mathbb{Z}_6$.

This presents us with a quandary. Do we require subrings to have the same unity element as their ambient ring, thus ruling out some interesting and otherwise valid cases, or do we relax this condition and just require a subring to have a unity element without worrying

too much about exactly which element fulfils this rôle? There is, again, no universally adopted convention, and after more than a century of intensive study of ring theory, it seems unlikely that a consensus will form any time soon.

In this book, we will adopt the conventions in the following definition, but you should be aware that many books use slightly different definitions or terms.

Definition 9.2 Let $R = (R, +, \cdot)$ be a ring, and let S be a subset of R . Then S is a **subring** of R if $(S, +_S, \cdot_S)$ is a ring in its own right, where $+_S$ and $\cdot_S$ are the restrictions to S of the addition and multiplication operations in R .

If S does not contain a unity element, then we say it is a **nonunital subring** or a **subring without unity**.

If the unity element 1_S in S is the same as the unity element 1_R in R , then we call S a **unital subring** or **subring with unity**.

In other words, unless otherwise specified we require a subring to have a unity element, but we don't necessarily require that unity element to be the same as the one from the larger ring. (We could perhaps refer to this situation as a **subring with different unity**, but that's a bit unwieldy and isn't in common usage.) We reserve the terms **unital subring** and **subring with unity** for the case where the two unity elements are the same. This is a somewhat complicated way of doing things, but seems to be the least worst option available.

The following proposition provides criteria for a subset $S \subseteq R$ to form a unital subring of R . Essentially, it says that we really just have to check whether S forms a subgroup of R under addition, is closed under multiplication, and contains the unity element $1 \in R$.

Proposition 9.3 Let $R = (R, +, \cdot)$ be a ring, and let $S \subseteq R$ be a subset of R . Then S is a subring of R if and only if the following conditions hold:

SR1 $(S, +)$ is a subgroup of $(R, +)$,

SR2 $a \cdot b \in S$ for all $a, b \in S$, and

SR3 there exists a unity element $1_S \in S$.

Furthermore, S is a unital subring of R if and only if

SR4 $1_S = 1_R$.

Proof This is all just a consequence of Definitions 8.1 and 9.2. Since $(R, +)$ is an abelian group, condition SR1 is equivalent to requiring that $(S, +)$ is an abelian group itself, which in turn is equivalent to condition R1 from Definition 8.1.

The closure condition SR2 is equivalent to requiring the restricted multiplication operation $\cdot_S$ to be well-defined on S . If it is, then it

will automatically be associative on S (since the full multiplication operation in R is associative) and hence condition R2 is satisfied. And since S is closed under multiplication and the distributive law R3 holds in R it must also hold in S . Next, SR3 ensures that R4 holds in S , and hence S is a subring of R . Finally, if SR4 is satisfied, then the unity element in S is the one inherited from R , and so S is a unital subring of R .

Conversely, if SR1 fails, then R1 also fails; if SR2 fails then the restricted multiplication operation $\cdot_S$ will not be well-defined; while if SR3 fails then S won't contain a unity element, and therefore can't be a subring of R . Finally, if condition SR4 isn't satisfied then S isn't a unital subring of R . $\square$

If only conditions SR1 and SR2 hold, then S is a nonunital subring of R .

For a subset $S \subseteq R$ to be a subring, we not only require S to satisfy the usual additive subgroup conditions (closure, existence of inverses) but also to satisfy certain multiplicative conditions (closure, existence of 1) as well. These extra requirements are more demanding, and as a result many rings which have entirely valid subgroups when considered just as additive abelian groups, fail to have corresponding subrings when we consider the ring structure as well. The following example illustrates this.

Example 9.4 The ring $\mathbb{Z}_4$ has no proper unital subrings.

The full ring $\mathbb{Z}_4$ and the trivial ring $\{0\}$ obviously form subrings, although the latter isn't unital in the sense of Definition 9.2. The only other candidate is the set $\{0, 2\}$, which certainly forms an additive subgroup of $\mathbb{Z}_4$. It's also closed under multiplication, but doesn't satisfy the last of the three conditions in Proposition 9.3, as it doesn't contain the unity element 1.

This is a special case of the following more general fact.

Proposition 9.5 *The rings $\mathbb{Z}$ and $\mathbb{Z}_n$ have no proper unital subrings.*

Proof Suppose that S is a nontrivial unital subring of $\mathbb{Z}_n$. Then it must contain the unity element $1 \in \mathbb{Z}_n$ (by SR3) and be closed under addition (by SR1). Together, these imply that S must also contain all multiples $k \cdot 1 = k$ of 1 for all $0 \leq k < n$, and therefore $S = \mathbb{Z}_n$.

Similarly, suppose that S is a unital subring of $\mathbb{Z}$. Then S must contain 1 and all finite multiples $k \cdot 1 = k$ for all $k \in \mathbb{Z}$, in which case $S = \mathbb{Z}$. $\square$

The reason that $\mathbb{Z}$ and $\mathbb{Z}_n$ have no proper unital subrings is that any such subring must contain the unity element. If we relax this

requirement then we have more possibilities: $\mathbb{Z}$ and $\mathbb{Z}_n$ have plenty of nonunital subrings:

Proposition 9.6 *For any $n \in \mathbb{N}$, the set*

$$n\mathbb{Z} = \{\dots, -2n, -n, 0, n, 2n, \dots\}$$

is a nonunital subring of $\mathbb{Z}$. Furthermore, all nonunital subrings of $\mathbb{Z}$ are of this form.

Proof For $n\mathbb{Z} \subseteq \mathbb{Z}$ to be a nonunital subring, we need only check conditions SR1 and SR2 from Proposition 9.3. The first of these, that $(n\mathbb{Z}, +)$ is an abelian subgroup of $(\mathbb{Z}, +)$, is straightforward, and follows from Proposition 2.7. Condition SR2 is also clear: for any $pn, qn \in n\mathbb{Z}$ we have $pn \cdot qn = pqn^2 = (pqn) \cdot n$, so $n\mathbb{Z}$ is closed under multiplication. Hence $n\mathbb{Z}$ is a nonunital subring of $\mathbb{Z}$.

Now suppose that $S \subseteq \mathbb{Z}$ is some nonunital subring of $\mathbb{Z}$. Then by SR1, $(S, +)$ has to be an abelian subgroup of $\mathbb{Z}$. But by Proposition 2.8 we know that this must be a cyclic subgroup, and hence $S = n\mathbb{Z}$ for some $n \in \mathbb{N}$ (the case $n = 0$ is ruled out by the requirement that S be nonunital). $\square$

Recall that the integer rings $\mathbb{Z}$ and $\mathbb{Z}_n$ were our motivating examples for the material in this and several subsequent chapters: it was by careful analysis of their essential properties that we arrived at the definition of a ring, and all of the other examples in this chapter were derived from them in one way or another. With this in mind, the following fact should not be so surprising.

Proposition 9.7 *Let R be a (unital) ring. Then R contains a subring isomorphic to $\mathbb{Z}$ or to $\mathbb{Z}_n$ for some $n > 1$.*

Proof The ring R is by definition closed under addition and multiplication, and contains a unity element 1. The subset

$$\langle 1 \rangle = \{\dots, -2, -1, 0, 1, 2, \dots\} \subseteq R$$

consists of everything in R generated by finite sums or products of 1. This is a subring of R : it is an abelian subgroup of R by Proposition 2.7, it is closed under multiplication, and it contains the unity element 1.

Recall from Proposition 8.19 that $\text{char}(R)$ is determined by the additive order of 1 in R . If $\text{char}(R) = 0$ then 1 has infinite additive order, and $\langle 1 \rangle \cong \mathbb{Z}$; if R has finite characteristic $\text{char}(R) = n$ then the additive order of 1 is also n , and hence $\langle 1 \rangle \cong \mathbb{Z}_n$. $\square$

The subring $\langle 1 \rangle$ is called the **prime subring**; it is the smallest nontrivial unital subring of R .

At the beginning of this section we saw that $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ has a subring isomorphic to $\mathbb{Z}_2$. This agrees with the above proposition, since

$\mathbb{Z}_2 \oplus \mathbb{Z}_2$ is a unital ring with characteristic 2. In general, the direct product $G \times H$ of two groups has at least one subgroup isomorphic to G and one isomorphic to H (see Proposition 3.29). Here, $G \times \{e\} \cong G$ and $\{e\} \times H \cong H$. Is this always the case for a direct sum $R \oplus S$?

A more general, if rather more technical, way of looking at this is to say that the direct sum operation is a **product** but not a **coproduct** in the category of unital rings.

As we learned in Example 4.16, this works for groups because there exist canonical **inclusion homomorphisms** $i_1: G \hookrightarrow G \times H$ and $i_2: H \hookrightarrow G \times H$ for any groups G and H . However, the corresponding inclusion functions $i_1: R \hookrightarrow R \oplus S$ and $i_2: S \hookrightarrow R \oplus S$ fail to be homomorphisms. The reason for this is that i_1 maps R 's unity element 1_R to $(1, 0)$ in $R \oplus S$ while i_2 maps 1_S to $(0, 1)$ in $R \oplus S$. But the unity element in $R \oplus S$ is $(1, 1)$ instead, and for i_1 and i_2 to be ring homomorphisms, we need them both to map unity elements to unity elements.

More generally, for $R \oplus S$ to have a subring isomorphic to R , there must exist an injective homomorphism $f: R \rightarrow R \oplus S$. If there is an injective homomorphism $g: R \rightarrow S$, then we can use this to construct the homomorphism $f = \text{id}_R \oplus g$ that maps $r \mapsto (r, g(r))$ for all $r \in R$. For example, $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ has a subring isomorphic to $\mathbb{Z}_2$ via the **diagonal** homomorphism $\Delta = \text{id}_R \oplus \text{id}_R$ defined by $\Delta(r) = (r, r)$ for all $r \in \mathbb{Z}_2$. While on the subject of direct sums of rings, it's a good time to state the following fact, which is an analogue of Proposition 1.32. This result is a form of the famous Chinese Remainder Theorem, which we will return to later in Chapter 10.

Proposition 9.8 (Chinese Remainder Theorem) *The ring $\mathbb{Z}_m \oplus \mathbb{Z}_n$ is isomorphic to the ring $\mathbb{Z}_{mn}$ if and only if m and n are coprime.*

Proof As before, we denote by $[a]_n$ the **residue** or **remainder** of a modulo n . Let $f: \mathbb{Z}_{mn} \rightarrow \mathbb{Z}_m \oplus \mathbb{Z}_n$ where $f(a) = ([a]_m, [a]_n)$. That is, we map any element of $\mathbb{Z}_{mn}$ to the ordered pair in $\mathbb{Z}_m \oplus \mathbb{Z}_n$ consisting of its residues modulo m and n , respectively.

This is certainly a ring homomorphism, since

$$\begin{aligned} f(a + b) &= ([a + b]_m, [a + b]_n) = ([a]_m + [b]_m, [a]_n + [b]_n) \\ &= ([a]_m, [a]_n) + ([b]_m, [b]_n) = f(a) + f(b) \end{aligned}$$

and

$$\begin{aligned} f(a \cdot b) &= ([ab]_m, [ab]_n) = ([a]_m [b]_m, [a]_n [b]_n) \\ &= ([a]_m, [a]_n) \cdot ([b]_m, [b]_n) = f(a) \cdot f(b). \end{aligned}$$

We need to show it's a bijection, which we can do by constructing a well-defined inverse homomorphism $f^{-1}: \mathbb{Z}_m \oplus \mathbb{Z}_n \rightarrow \mathbb{Z}_{mn}$.

If m and n are coprime then Euclid's Algorithm enables us to find integers $p, q \in \mathbb{Z}$ such that $pm + qn = 1$. We can use this to define

$$f^{-1}(a, b) = [pmb + qna]_{mn}$$

for all $a \in \mathbb{Z}_m$ and $b \in \mathbb{Z}_n$. To check that this is indeed the inverse of f , we need to verify that $f(f^{-1}(a, b)) = (a, b)$ and $f^{-1}(f(k)) = k$ for all $a \in \mathbb{Z}_m$, $b \in \mathbb{Z}_n$ and $k \in \mathbb{Z}_{mn}$. Firstly,

$$\begin{aligned} f(f^{-1}(a, b)) &= f([pmb + qna]_{mn}) \\ &= ([pmb + qna]_{mn})_m, ([pmb + qna]_{mn})_n \\ &= ([pmb + qna]_m, [pmb + qna]_n) \\ &= ([qna]_m, [pmb]_n). \end{aligned}$$

But $pm + qn = 1$ and hence $pm = 1 - qn$, so $pm \equiv 1 \pmod{n}$, therefore $pmb \equiv b \pmod{n}$. Similarly, $qn = 1 - pm$, so $qn \equiv 1 \pmod{m}$, hence $qna \equiv a \pmod{m}$. Thus $([qna]_m, [pmb]_n) = (a, b)$. Secondly,

$$\begin{aligned} f^{-1}(f(k)) &= f^{-1}([k]_{mn}) \\ &= [pm[k]_n + qn[k]_m]_{mn} \\ &= [pmk + qnk]_{mn} \\ &= [(pm + qn)]_{mn}[k]_{mn} \\ &= k. \end{aligned}$$

Therefore f is an isomorphism.

If, on the other hand, m and n are not coprime, then by Proposition 1.32 the additive abelian groups $(\mathbb{Z}_{mn}, +)$ and $(\mathbb{Z}_m \oplus \mathbb{Z}_n, +)$ can't be isomorphic since the order of $(1, 1)$ in the latter group is equal to $\text{lcm}(m, n)$ which is strictly less than mn , the order of 1 in $\mathbb{Z}_{mn}$. $\square$

Corollary 9.9 Suppose that $n = p_1^{m_1} \dots p_k^{m_k}$, where $p_1, \dots, p_k$ are distinct primes, and $m_1, \dots, m_k \in \mathbb{N}$. Then

$$\mathbb{Z}_n \cong \mathbb{Z}_{p_1^{m_1}} \oplus \dots \oplus \mathbb{Z}_{p_k^{m_k}}.$$

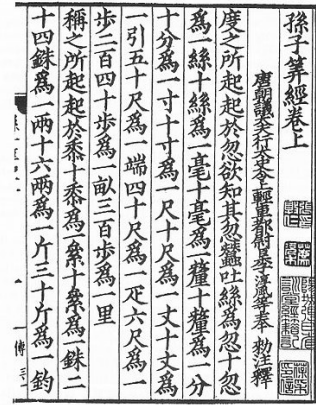
Proof If $p_1, \dots, p_k$ are distinct primes, then $p_1^{m_1}, \dots, p_k^{m_k}$ are pairwise coprime. The result then follows by induction on k . $\square$

Back to our discussion of subrings. What about the other examples of rings we've met so far? Do polynomial rings or matrix rings have interesting subrings? Well, as it happens, we've met some matrix subrings already:

Example 9.10 For any ring R , the matrix rings $UT_n(R)$, $LT_n(R)$ and $D_n(R)$ of, respectively, upper-triangular, lower-triangular and diagonal matrices, are all subrings of $M_n(R)$.

Furthermore, $D_n(R)$ is a subring of both $UT_n(R)$ and $LT_n(R)$, and all of them contain the prime subring $\langle I \rangle \cong \mathbb{Z}$.

This example raises another interesting question: the ring $D_n(R)$ of diagonal matrices is a subring of both $UT_n(R)$ and $LT_n(R)$; in fact it is exactly the intersection $UT_n(R) \cap LT_n(R)$. Proposition 2.9 tells us



Wikimedia Commons

The earliest known statement of the **Chinese Remainder Theorem** occurs in the book *Sun Tzu Suan Ching* (or *Sūn Zǐ Suàn Jīng*), which translates roughly as *The Mathematical Classic of Sun Tzu*. This book, a page of which is shown above, dates from the third century AD, and is one of a series of important mathematical texts known during the Tang dynasty (618–907AD) as the “Ten Computational Canons”.

Its author **Sun Tzu** (or **Sūn Zǐ**) is believed to be distinct from the military strategist of the same name, the author of *The Art of War*, who is thought to have lived during the sixth century BC. A more detailed discussion of the problem appears in the book *Āryabhaṭīya* (499AD) by the Indian mathematician **Āryabhaṭa** (476–550AD), and also in the *Liber Abaci* (1202) of **Fibonacci** (c.1170–c.1250).

The number-theoretic version of the theorem is as follows:

Chinese Remainder Theorem Suppose that $n_1, \dots, n_k \in \mathbb{N}$ are pairwise coprime. Then for any $a_1, \dots, a_k \in \mathbb{Z}$ there exists $x \in \mathbb{Z}$ such that

$$\begin{aligned} x &\equiv a_1 \pmod{n_1}, \\ &\vdots \\ x &\equiv a_k \pmod{n_k}. \end{aligned}$$

Furthermore, any two solutions x and y are congruent modulo $n = n_1 \dots n_k$.

The surjectivity of the homomorphism f in Proposition 9.8 ensures the existence of x , and its injectivity ensures the modulo- n congruence property.

that the intersection of two subgroups is itself a subgroup, so does the same thing happen for subrings?

Proposition 9.11 *Let $S, T \subseteq R$ be unital subrings of some ring R . Then their intersection $S \cap T$ is also a unital subring of R .*

Proof Condition SR₁ follows from Proposition 2.9. To check the closure requirement SR₂, we observe that for any two elements $a, b \in S \cap T$, their product $a \cdot b$ must lie in S (because S is a subring and thus closed under multiplication) and also in T (because T is also a subring and hence also closed under multiplication). Since $a \cdot b \in S$ and $a \cdot b \in T$, it follows that $a \cdot b \in S \cap T$. Finally, SR₃ holds because both S and T are themselves unital subrings and must therefore contain the unity element $1 \in R$, and so their intersection does as well. $\square$

The matrix ring $M_n(\mathbb{R})$ has subrings isomorphic to $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and, somewhat unexpectedly, $\mathbb{C}$:

Example 9.12 The matrix ring $M_n(\mathbb{R})$ has characteristic 0, since the identity matrix I has infinite additive order. By Proposition 9.7, the prime subring

$$\langle I \rangle = \{nI : n \in \mathbb{Z}\}$$

must therefore be isomorphic to $\mathbb{Z}$. We can extend this subring to get the subrings

$$\{qI : q \in \mathbb{Q}\} \cong \mathbb{Q} \quad \text{and} \quad \{xI : x \in \mathbb{R}\} \cong \mathbb{R}.$$

The subring

$$C = \left\{ \begin{bmatrix} a & b \\ -b & a \end{bmatrix} : a, b \in \mathbb{R} \right\}$$

of $M_2(\mathbb{R})$ is isomorphic to the ring $\mathbb{C}$ via the isomorphism

$$f: \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mapsto a + bi.$$

With a little bit of work, we can extend this to a subring of $M_n(\mathbb{R})$ for $n \geq 2$, by considering the subset of $n \times n$ matrices of the form

$$\begin{bmatrix} a & b & 0 & \cdots & 0 \\ -b & a & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

for $a, b \in \mathbb{R}$.

There are other simple ways of constructing subrings of $M_n(R)$. For example, if S is a subring of R , then $M_n(S)$ will be a subring of $M_n(R)$. Or, for $0 < k < n$, the ring $M_k(R)$ embeds neatly as a subring of

$M_n(R)$: map a given $k \times k$ matrix A to the $n \times n$ matrix whose top left $k \times k$ submatrix is a copy of A , whose other diagonal entries are 1, and has zero everywhere else.

9.2 Homomorphisms and ideals

RECALL THAT a function $f: R \rightarrow S$ is a **(ring) homomorphism** if

$$f(a + b) = f(a) + f(b) \quad \text{and} \quad f(a \cdot b) = f(a) \cdot f(b)$$

for all $a, b \in R$.¹

Example 9.13 For a ring R , the **identity homomorphism** $\text{id}_R: R \rightarrow R$ maps $a \mapsto a$ for all $a \in R$. This homomorphism is an isomorphism.

Example 9.14 For any rings R and S , the **zero homomorphism** $z: R \rightarrow S$ maps $a \mapsto 0$ for all $a \in R$.

Example 9.15 Let $f: \mathbb{C} \rightarrow \mathbb{C}$ by $z \mapsto \bar{z}$. Then

$$\begin{aligned} f(z + w) &= \overline{z + w} = \bar{z} + \bar{w} = f(z) + f(w), \\ f(zw) &= \overline{zw} = \bar{z}\bar{w} = f(z)f(w). \end{aligned}$$

and so complex conjugation is a ring homomorphism. Since it's also bijective, it's an automorphism too.

Back in Chapter 4 we obtained some basic properties satisfied by group homomorphisms. In particular, we found that $f(e_G) = e_H$ for any group homomorphism $f: G \rightarrow H$;² that is, a group homomorphism maps the identity of its domain to the identity of its codomain. Proposition 8.16 confirmed that ring homomorphisms map additive identities to additive identities, and also that *surjective* ring homomorphisms map unity elements to unity elements (at least for rings that have such elements).

But Proposition 8.16 has a loophole: what happens to the unity element when the homomorphism isn't surjective?

Proposition 9.16 Let $f: R \rightarrow S$ be a ring homomorphism. Then

$$f(1_R)^2 = f(1_R).$$

Proof Observe that $f(1_R) = f(1_R \cdot 1_R) = f(1_R) \cdot f(1_R)$. □

What this means is that a non-surjective ring homomorphism needn't map the unity element $1_R \in R$ to the unity element $1_S \in S$, although certainly this is permitted and will often be the case with many of the rings and homomorphisms we'll consider. However, all we really require is that 1_R is mapped to some element $s \in S$ satisfying the

Mrs Erlynne: Ideals are dangerous things. Realities are better. They wound, but they're better.

Lady Windermere: If I lost my ideals, I should lose everything.

— Oscar Wilde (1854–1900),
Lady Windermere's Fan (1893)

¹ Definition 8.14, page 323.

² Proposition 4.24, page 110.

condition $s^2 = s$.

Such an element is said to be **idempotent**. It is always the case that the zero element and (if it exists) the unity element are idempotent. Sometimes, however, there may be other elements satisfying this condition as well.

This leads us to the following clarification.

Definition 9.17 Let $f: R \rightarrow S$ be a ring homomorphism. Then f is **unital** if $f(1_R) = 1_S$, and **nonunital** if $f(1_R) \neq 1_S$, or if either R or S are nonunital rings.

Any nontrivial ring R admits at least one unital homomorphism, namely the identity map $\text{id}_R: R \rightarrow R$ mapping $a \mapsto a$ for all $a \in R$, and at least one nonunital homomorphism, the zero map $z_R: R \rightarrow R$ given by $a \mapsto 0$ for all $a \in R$.

³ This is a consequence of the fact that $\mathbb{Z}$ is an integral domain.

The integer ring $\mathbb{Z}$ has only two idempotent elements, 0 and 1.³ Because of this, the zero homomorphism is the only nonunital homomorphism from $\mathbb{Z}$ to itself.

In fact, an endomorphism $f: \mathbb{Z} \rightarrow \mathbb{Z}$ is completely determined by what happens to 1. We know from Proposition 9.16 above that $f(0) = 0$, and 1 is very nearly as limited in its choice of destination: it can map either to 0 or 1. The former gives us the zero homomorphism while the latter gives the identity homomorphism. The destination of any other $n \in \mathbb{Z}$ is determined by

$$f(n) = f(n \cdot 1) = f(1 + \cdots + 1) = f(1) + \cdots + f(1) = n \cdot f(1).$$

In general, the image of the prime subring $\langle 1 \rangle$ will be completely determined in this way.

We will usually require ring homomorphisms to be unital. Nonunital homomorphisms are interesting, but unital ones have certain nice properties that will sometimes come in useful. In particular, as we'll see in Chapter 10, they map invertible elements to invertible elements. The following two examples show that nonunital homomorphisms can often be defined between relatively straightforward unital rings.

Example 9.18 Let $f: \mathbb{Z}_2 \rightarrow \mathbb{Z}_6$ such that $0 \mapsto 0$ and $1 \mapsto 3$, and let $g: \mathbb{Z}_3 \rightarrow \mathbb{Z}_6$ such that $0 \mapsto 0$, $1 \mapsto 4$ and $2 \mapsto 2$. Then f and g are both nonunital homomorphisms.

Example 9.19 Let $f: \mathbb{Z} \rightarrow M_2(\mathbb{Z})$ by $n \mapsto \begin{bmatrix} 2n & n \\ -2n & -n \end{bmatrix}$. This satisfies the homomorphism conditions, but doesn't map $1 \in \mathbb{Z}$ to the 2×2 identity matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \in M_2(\mathbb{Z})$.

An important class of homomorphisms can be defined on polynomial rings:

Example 9.20 Let $\mathbb{Z}[x]$ denote the ring of polynomials with integer coefficients. Then we can regard an arbitrary polynomial $p = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}[x]$ as a function $p: \mathbb{Z} \rightarrow \mathbb{Z}$. (In fact, this is usually the first context in which we meet polynomials.)

So, for example, the polynomial $q = x^2 - 2x + 1$ defines a function $q: \mathbb{Z} \rightarrow \mathbb{Z}$ which amongst other things maps $1 \mapsto 1^2 - 2 \cdot 1 + 1 = 0$, maps $0 \mapsto 0^2 - 2 \cdot 0 + 1 = 1$, and maps $-1 \mapsto (-1)^2 - 2(-1) + 1 = 4$, and so on.

A different way of looking at this is to keep the argument fixed and vary the polynomial. So, for some other polynomial $r = 5x^4 + 2x - 1$ we have $r(1) = 5 \cdot 1^4 + 2 \cdot 1 - 1 = 6$. This viewpoint enables us to define a function $f_1: \mathbb{Z}[x] \rightarrow \mathbb{Z}$ given by $p \mapsto p(1) \in \mathbb{Z}$ for any polynomial $p \in \mathbb{Z}[x]$.

So, in particular, $f_1(q) = q(1) = 0$ while $f_1(r) = r(1) = 4$, and so forth.

More generally, for any $n \in \mathbb{Z}$ we can define a function $\text{ev}_n: \mathbb{Z}[x] \rightarrow \mathbb{Z}$ in the analogous way, mapping a given polynomial p to its **evaluation** $p(n) \in \mathbb{Z}$.

Even more generally, for any ring R and fixed element $a \in R$ we can define the **evaluation homomorphism** $\text{ev}_a: R[x] \rightarrow R$ such that $p \mapsto p(a) \in R$.

This can further be extended to multivariate polynomial rings: given a ring R and an ordered k -tuple $(a_1, \dots, a_k)$ of elements of R , we can define the evaluation homomorphism $\text{ev}_{(a_1, \dots, a_k)}: R[x_1, \dots, x_k] \rightarrow R$ by $p \mapsto p(a_1, \dots, a_k) \in R$.

Now we look at some of the constructions that proved so important in group theory. In particular we want to define images and kernels for ring homomorphisms, and also derive the correct notion of a coset. The first of these is straightforward: if $f: R \rightarrow S$ is a (possibly nonunital) homomorphism of (possibly nonunital) rings R and S , then its image is defined as

$$\text{im}(f) = \{s \in S : s = f(r) \text{ for some } r \in R\}.$$

Proposition 4.25 tells us that the image of a group homomorphism is a subgroup of its codomain. Does the same thing happen for ring homomorphisms? Is the image of a ring homomorphism always a subring of its codomain? We have to be a little careful here, and consider the various possibilities.

Proposition 9.21 *Let R and S be rings, with or without unity elements, and let $f: R \rightarrow S$ be a (possibly nonunital) ring homomorphism. Then:*

- (i) *If R is nonunital, then $\text{im } f$ is a (possibly nonunital) subring of S .*

(ii) *If R is unital then $\text{im } f$ is a subring of S , and if f is a unital homomorphism, $\text{im } f$ is a unital subring of S .*

Proof Proposition 4.25 ensures that condition SR1 is satisfied: $\text{im } f$ is an abelian subgroup of S . Closure under multiplication, condition SR2, is satisfied since for any two elements $f(a), f(b) \in \text{im } f$ the fact that f is a ring homomorphism ensures that

$$f(a) \cdot f(b) = f(a \cdot b).$$

Hence $\text{im } f$ is a (possibly nonunital) subring of S .

If R has a unity element 1_R then for any $f(a) \in \text{im } f$ we have

$$f(1_R) \cdot f(a) = f(1_R \cdot a) = f(a),$$

and so $f(1_R)$ acts as a unity element in $\text{im } f$. This satisfies condition SR3, and so $\text{im } f$ is a subring of S .

If, in addition, f is a unital homomorphism, we know that $f(1_R) = 1_S$, which is unique by Proposition 8.9, and since we saw just now that $f(1_R)$ is the unity element in $\text{im } f$, condition SR4 is satisfied, and $\text{im } f$ is a unital subring of S . $\square$

Example 9.22 Let $f: \mathbb{Z}_2 \rightarrow \mathbb{Z}_6$ be the homomorphism from Example 9.18 mapping $0 \mapsto 0$ and $1 \mapsto 3$. Then $\text{im}(f) = \{0, 3\} \cong \mathbb{Z}_2$ is a subring of $\mathbb{Z}_6$, but not a unital subring since f isn't a unital homomorphism.

Similarly, let $g: \mathbb{Z}_3 \rightarrow \mathbb{Z}_6$ be the other homomorphism from Example 9.18 mapping $0 \mapsto 0$, $1 \mapsto 4$ and $2 \mapsto 2$. Then $\text{im}(g) = \{0, 2, 4\} \cong \mathbb{Z}_3$ is also a subring of $\mathbb{Z}_6$ but again not a unital subring.

Example 9.23 Let $f: \mathbb{Z} \rightarrow M_2(\mathbb{Z})$ be the nonunital homomorphism from Example 9.19 mapping $n \mapsto \begin{bmatrix} 2n & n \\ -2n & -n \end{bmatrix}$. Then $\text{im}(f)$ is a subring of $M_2(\mathbb{Z})$ isomorphic to $\mathbb{Z}$. It has unity element $f(1) = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix}$ but is not a unital subring in the sense of Definition 9.2.

The definition of the image of a ring homomorphism is straightforward: it's the obvious analogue of the corresponding notion for functions and group homomorphisms. Less obvious is how we should go about defining kernels of ring homomorphisms.

For group homomorphisms, we define the kernel as the preimage of the identity element; that is, everything in the domain that maps to the identity element in the codomain. But a ring has two elements that fulfil that rôle, and it's not immediately obvious which one we should use.

In fact, there are a few reasons why we should use the additive identity 0 rather than the multiplicative identity 1. Firstly, we want our definition to extend to nonunital rings as well as unital ones,

and using the multiplicative identity would prevent us from doing that. Secondly, rings are in some sense abelian groups with extra structure, and ring homomorphisms are special kinds of abelian group homomorphisms, so we might reasonably expect ring homomorphism kernels to be special sorts of abelian group homomorphisms. Thirdly, as we'll see in a little while, using the additive identity enables us to define the correct analogues of cosets and normal subgroups so that we can formulate a nice theory of quotient rings.

Definition 9.24 Let $f: R \rightarrow S$ be a ring homomorphism. Then the **kernel** of f is the set

$$\ker(f) = \{r \in R : f(r) = 0_s\}.$$

In Chapter 4 we found that kernels of group homomorphisms are normal subgroups (and also that every normal subgroup is the kernel of some group homomorphism).⁴ For kernels of ring homomorphisms, things are slightly more complicated. ⁴ Proposition 4.29, page 113.

If $f: R \rightarrow S$ is a ring homomorphism then certainly $\ker(f)$ is an additive (normal) subgroup of $(R, +)$ by Proposition 4.29. And $\ker(f)$ is closed under multiplication, since for any $a, b \in \ker(f)$ we have $f(a) = f(b) = 0$, and so

$$f(a \cdot b) = f(a) \cdot f(b) = 0 \cdot 0 = 0,$$

hence $a \cdot b \in \ker(f)$ as well. So kernels of ring homomorphisms are at least nonunital subrings. But $\ker(f)$ need not have a unity element, so it need not qualify for our definition of a subring.⁵ And even if it does, that unity element won't necessarily be the same as the unity element in the ambient ring. As we'll see in a little while, $\ker(f)$ will be a unital subring in very specific circumstances. ⁵ Definition 9.2, page 345.

For the moment, however, we'll look at a few examples.

Example 9.25 Let $f: \mathbb{Z}_6 \rightarrow \mathbb{Z}_2$ and $g: \mathbb{Z}_6 \rightarrow \mathbb{Z}_3$ such that

$$f(k) = \begin{cases} 0 & \text{if } k = 0, 2, 4 \\ 1 & \text{if } k = 1, 3, 5 \end{cases} \quad \text{and} \quad g(k) = \begin{cases} 0 & \text{if } k = 0, 3 \\ 1 & \text{if } k = 1, 4 \\ 2 & \text{if } k = 2, 5 \end{cases}$$

Then $\ker(f)$ and $\ker(g)$ are both subrings of $\mathbb{Z}_6$. Each has a unity element (4 in the case of $\ker(f)$, and 3 in $\ker(g)$) but neither is equal to 1, the unity element of the ambient ring $\mathbb{Z}_6$.

Example 9.26 Let n be a positive integer greater than 1. Then we can define a homomorphism $f: \mathbb{Z} \rightarrow \mathbb{Z}_n$ such that $f(k) = [k]_n$ for all $k \in \mathbb{Z}$. Then $\ker(f) = n\mathbb{Z}$, which we know from Proposition 9.6 is a nonunital subring of $\mathbb{Z}$.

Example 9.27 Let $a \in \mathbb{Z}$ and let $\text{ev}_a: \mathbb{Z}[x] \rightarrow \mathbb{Z}$ be the evaluation homomorphism from Example 9.20. Then $\ker(\text{ev}_a)$ consists of those polynomials p in $\mathbb{Z}[x]$ for which $p(a) = 0$. In other words, $\ker(\text{ev}_a)$ comprises all polynomials which have a as a root, or equivalently all polynomials which have $(x-a)$ as a factor.

This kernel is a nonunital subring of $\mathbb{Z}[x]$: it doesn't contain $1 \in \mathbb{Z}[x]$, and no other element serves as a unity element.

Example 9.26 shows that the kernel of the modulo- n residue map $f_n = [-]_n: \mathbb{Z} \rightarrow \mathbb{Z}_n$ is the nonunital subring $n\mathbb{Z} \subseteq \mathbb{Z}$ consisting of all integer multiples of n . Any multiple of n times an arbitrary integer is also a multiple of n , since $kn \cdot m = (km)n \in n\mathbb{Z}$.

Similarly, Example 9.27 shows that the kernel of the evaluation homomorphism $\text{ev}_a: \mathbb{Z}[x] \rightarrow \mathbb{Z}$ consists of all polynomials p that have $(x-a)$ as a factor (or, equivalently, for which $p(a) = 0$). Any other polynomial $q \in \mathbb{Z}[x]$ multiplied by one of these polynomials will also have $(x-a)$ as a factor, and therefore belong to $\ker(\text{ev}_a)$.

More precisely, for any $k \in \ker(f_n)$ and any $m \in \mathbb{Z}$ both $m \cdot k$ and $k \cdot m$ lie in $\ker(f_n)$. Similarly, for any polynomial $p \in \ker(\text{ev}_a)$ and any polynomial $q \in \mathbb{Z}[x]$ both $p \cdot q$ and $q \cdot p$ lie in $\ker(\text{ev}_a)$.

In fact, this is true for the kernel of any ring homomorphism:

Proposition 9.28 *Let $f: R \rightarrow S$ be a ring homomorphism. Then for any $k \in \ker(f)$ and $a \in R$ we have $k \cdot a \in \ker(f)$ and $a \cdot k \in \ker(f)$.*

Proof This follows almost immediately from the multiplicative property of a ring homomorphism, and from Proposition 8.8 (i), since

$$f(a \cdot k) = f(a) \cdot f(k) = f(a) \cdot 0 = 0,$$

so $a \cdot k \in \ker(f)$. A similar argument confirms $k \cdot a \in \ker(f)$ too. $\square$

Motivated by this discussion and by Proposition 4.29 we are ready to introduce the following definition.

Definition 9.29 Let R be a ring and I be a subset of R satisfying the first two subring conditions from Definition 9.2:

I1 $(I, +)$ is a subgroup of $(R, +)$, and

I2 $a \cdot b \in I$ for all $a, b \in I$.

Then I is a **left ideal** of R if

I3L $a \cdot i \in I$ for all $i \in I$ and $a \in R$,

a **right ideal** of R if

I3R $i \cdot a \in I$ for all $i \in I$ and $a \in R$,

and an **ideal** (or sometimes a **two-sided ideal**) if both I3L and I3R (the **absorption conditions**) are satisfied.

Ideals are the ring-theoretic analogues of normal subgroups, and in the next section we will learn how to use them to form quotient rings. In particular, from Proposition 9.28 and the discussion following Definition 9.24 we know that kernels of ring homomorphisms are ideals. That is, for any ring homomorphism $f: R \rightarrow S$, the kernel $\ker(f)$ is an ideal of R . Later on, we'll show the converse, that all ideals can be regarded as kernels of ring homomorphisms. For the moment, however, we will look at a few more examples.

Example 9.30 Let R be a commutative unital ring and let $a \in R$. Then

$$\langle a \rangle = \{ra : r \in R\}$$

is the **principal ideal** of R generated by a .

If R is a noncommutative unital ring, then for any $a \in R$ we can define the **left** and **right principal ideals**

$$aR = \{ar : r \in R\} \quad \text{and} \quad Ra = \{ra : r \in R\}$$

and the **(two-sided) principal ideal**

$$RaR = \left\{ \sum_{i=1}^n r_i a s_i : r_i, s_i \in R, n \in \mathbb{N} \right\}$$

generated by a .

A (left, right or two-sided) principal ideal is the largest ideal that can be generated from a single given element of the ring in question.

Example 9.31 The previous example can be extended to finite sets of elements of a commutative unital ring R . Suppose that $a_1, \dots, a_k \in R$. Then

$$\langle a_1, \dots, a_k \rangle = \{r_1 a_1 + \dots + r_k a_k : r_1, \dots, r_k \in R\}$$

is the ideal of R generated by $a_1, \dots, a_k$.

The following are two concrete examples of principal ideals.

Example 9.32 The principal ideal $\langle x \rangle$ of $\mathbb{Z}[x]$ generated by x consists of those polynomials in $\mathbb{Z}[x]$ that have zero constant term.

This is clearly an ideal, since it forms an additive abelian group, is closed under multiplication, and satisfies both the left and right absorption conditions. Also, it is equal to the kernel of the evaluation homomorphism $\text{ev}_0: \mathbb{Z}[x] \rightarrow \mathbb{Z}$, which is known to be an ideal.

Example 9.33 The principal ideal $\langle n \rangle$ of $\mathbb{Z}$ generated by a given positive integer n consists of all integer multiples of n . This is exactly the ideal discussed in Example 9.26.

Every ring R has at least two ideals:

Example 9.34 For any ring R , the trivial subring $\{0\}$ is an ideal: it satisfies the left and right absorption conditions by Proposition 8.8 (i).

Example 9.35 Any ring R is an ideal of itself, with the left and right absorption conditions reducing to the requirement that R be closed under multiplication, which it is.

Definition 9.29 also includes two slightly more general objects: left and right ideals. If a ring R is commutative, then clearly any left ideal will also be a right ideal, and vice versa. However, if R isn't commutative then we have the possibility of left ideals that aren't right ideals, and right ideals that aren't left ideals. The following example explores this idea.

Example 9.36 Let

$$I = \left\{ \begin{bmatrix} 0 & x \\ 0 & y \end{bmatrix} : x, y \in \mathbb{R} \right\}.$$

This is a left ideal but not a right ideal of the ring $M_2(\mathbb{R})$, since it forms an abelian group under addition, is closed under addition and multiplication, and satisfies the left absorption condition I_3L from Definition 9.29. Similarly, the set

$$J = \left\{ \begin{bmatrix} 0 & 0 \\ x & y \end{bmatrix} : x, y \in \mathbb{R} \right\}$$

is a right ideal but not a left ideal in $M_2(\mathbb{R})$.

In most of what follows, however, we will focus on two-sided ideals, because they are exactly what we need to define quotient rings.

Example 9.37 The ring $\mathbb{Z}_{12}$ has six ideals, namely

$$\begin{aligned} \langle 0 \rangle &= \{0\}, & \langle 6 \rangle &= \{0, 6\}, & \langle 4 \rangle &= \{0, 4, 8\}, \\ \langle 3 \rangle &= \{0, 3, 6, 9\}, & \langle 2 \rangle &= \{0, 2, 4, 6, 8, 10\}, & \langle 1 \rangle &= \mathbb{Z}_{12}. \end{aligned}$$

We can arrange them into the lattice diagram shown in Figure 9.1.

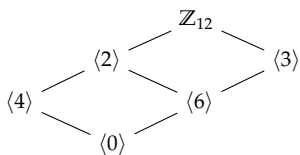


Figure 9.1: Lattice diagram for the ring $\mathbb{Z}_{12}$

The only one of these ideals of $\mathbb{Z}_{12}$ that contains the unity element 1 is the full ring $\mathbb{Z}_{12} = \langle 1 \rangle$. Furthermore, the only ideal of $M_2(\mathbb{Z})$ that contains the identity $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is $M_2(\mathbb{Z})$ itself, and $\mathbb{R}[x]$ is the only ideal of $\mathbb{R}[x]$ that contains 1. This is true in general:

Proposition 9.38 Let R be a unital ring, and let I be an ideal of R that contains 1. Then $I = R$.

Proof The left and right absorption conditions imply that $i \cdot r \in I$ and $r \cdot i \in I$ for any $i \in I$ and $r \in R$. In particular, setting $i = 1$ it follows that $1 \cdot r = r = r \cdot 1 \in I$ for any $r \in R$, and hence $R \subseteq I$. Since $I \subseteq R$ by definition, we have $I = R$. $\square$

We can generalise this result further, to show that if an ideal contains

any invertible element then it must be the entire ring.

Proposition 9.39 *Let R be a unital ring, and let I be an ideal of R that contains a unit u . Then $I = R$.*

Proof Suppose that $u \in I$ is a unit. Then the left and right absorption conditions imply that $u \cdot r \in I$ and $r \cdot u \in I$ for any $r \in R$. Setting $r = u^{-1}$ we see in particular that $u \cdot u^{-1} = 1 = u^{-1} \cdot u \in I$ as well, and hence by Proposition 9.38 it follows that $I = R$. $\square$

9.3 Quotient rings

NOW WE HAVE what seems to be the correct notion of a “normal subring”, namely an **ideal**, we have to devise the corresponding ring-theoretic definition of cosets. Again, we must decide whether to adopt the additive approach and use cosets of the form $a+I$, or to use multiplicative cosets of the form aI . As before, taking the view that rings are special types of additive abelian groups, we follow the additive approach to get the following definition.

Definition 9.40 Let R be a ring, and let I be an ideal of R . Then for any element $a \in R$ we define the **coset**

$$a+I = \{a+i : i \in I\}.$$

A little further thought gives us another reason why multiplicative cosets are unsuitable for our purposes here. We need the cosets to partition the full ring into subsets of equal size, and if we have a proper ideal $I \subset R$, the absorption conditions I_3L and I_3R from Definition 9.29 ensure that for all elements $a \in R$ and $i \in I$, both $a \cdot i$ and $i \cdot a$ lie in I . So no element of $R \setminus I$ belongs to any multiplicative coset aI or Ia .

Furthermore, Proposition 4.38 tells us that for some homomorphism $f: R \rightarrow S$ with $I = \ker(f)$, the additive cosets $a+I$ can be expressed as preimages of f in the same way as for group homomorphisms.

Now that we have the correct ring-theoretic versions of normal subgroups and cosets, we want to use them to define quotient rings. The approach we will use is the obvious analogue of the one we used for quotient groups. Given an ideal $I \subseteq R$, we define R/I to be the set of cosets of I in R . We can then define a suitable addition operation in the same way as for quotient groups, at which point it doesn't seem too unreasonable to ask if we can also define an appropriate multiplication operation as well. The following proposition confirms this and makes the construction precise.

My methods are methods of working and thinking, and have therefore anonymously invaded everywhere.

— Emmy Noether (1882–1935),
letter to Helmut Hasse (1898–1979),
12 November 1931

Proposition 9.41 *Let R be a (possibly nonunital) ring and let $I \subseteq R$ be an ideal. Then the set*

$$R/I = \{a+I : a \in R\}$$

*forms a (possibly nonunital) **quotient ring**, with addition and multiplication operations*

$$(a+I) + (b+I) = (a+b)+I$$

$$(a+I) \cdot (b+I) = (a \cdot b)+I$$

for any $a, b \in R$. Furthermore, if R is unital, so is R/I , and if R is commutative, R/I is as well.

Proof First we must prove that R/I forms an abelian group with the specified addition operation, and this essentially follows from Proposition 3.24. The addition operation is well-defined and associative, the additive identity is the coset $I = 0+I$, and any coset $a+I$ has inverse $-a+I$.

The multiplication operation in R/I is associative since the one in R is: for any $a, b, c \in R$ we have

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

The distributivity conditions are satisfied too, since

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

and

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

If R is unital then so is R/I : the coset $1+I$ acts as a unity element in R/I since

$$(1+I) \cdot (a+I) = (1 \cdot a)+I = a+I = (a \cdot 1)+I = (a+I) \cdot (1+I)$$

for any $a \in R$. And if R is commutative, then for any $a, b \in R$ we have

$$(a+I) \cdot (b+I) = (a \cdot b)+I = (b \cdot a)+I = (b+I) \cdot (a+I),$$

and hence R/I is commutative. $\square$

Having now introduced the general construction and proved that it works, it's time to look at some examples. The first one is very similar to our motivating example of quotient groups from Section 3.2.

Example 9.42 Let $f_n: \mathbb{Z} \rightarrow \mathbb{Z}_n$ be the ring homomorphism that maps an integer k to its modulo- n equivalence class $[k]_n$. Then the kernel $K = \ker(f_n)$ is an ideal of $\mathbb{Z}$ and we can therefore form the quotient ring $\mathbb{Z}/K$.

The kernel K consists of all integers congruent to 0 modulo n ; that is, all integers of the form kn . A coset $m+K$ therefore consists of integers of the form $kn+m$; that is, exactly those integers that are congruent to m modulo n . There are thus n of these, namely $K = 0+K, 1+K, \dots, (n-1)+K$.

The quotient ring $\mathbb{Z}/K$ therefore consists of these n cosets. Its additive structure is exactly addition modulo n , and its multiplicative structure is multiplication modulo n . Hence $\mathbb{Z}/K \cong \mathbb{Z}_n$.

This example should look familiar. We have a surjective homomorphism $f_n: \mathbb{Z} \rightarrow \mathbb{Z}_n$, which means that $\text{im}(f_n) = \mathbb{Z}_n$, and the example above shows that $\mathbb{Z}/\ker(f_n) \cong \mathbb{Z}_n$. Putting all this together we have $\mathbb{Z}/\ker(f_n) \cong \text{im}(f_n)$, which is a ring-theoretic example of the First Isomorphism Theorem for groups.⁶ We'll come back to this in a little while when we state and prove ring-theoretic versions of all three Isomorphism Theorems.

⁶ Theorem 4.40, page 117.

Example 9.43 Looking again at the homomorphisms $f: \mathbb{Z}_6 \rightarrow \mathbb{Z}_2$ and $g: \mathbb{Z}_6 \rightarrow \mathbb{Z}_3$ in Example 9.25, we have

$$K_1 = \ker(f) = \{0, 2, 4\} \quad \text{and} \quad K_2 = \ker(g) = \{0, 3\}.$$

The cosets of K_1 in $\mathbb{Z}_6$ are

$$\begin{aligned} K_1 &= 0+K_1 = 2+K_1 = 4+K_1 = \{0, 2, 4\} \\ \text{and} \quad 1+K_1 &= 3+K_1 = 5+K_1 = \{1, 3, 5\}. \end{aligned}$$

The quotient ring $\mathbb{Z}_6/K_1$ thus has two elements and must therefore be isomorphic to $\mathbb{Z}_2$. By examining the addition and multiplication of these cosets, we find that this is indeed the case.

The cosets of K_2 in $\mathbb{Z}_6$ are

$$\begin{aligned} K_2 &= 0+K_2 = 3+K_2 = \{0, 3\}, \\ 1+K_2 &= 4+K_2 = \{1, 4\}, \\ \text{and} \quad 2+K_2 &= 5+K_2 = \{2, 5\}. \end{aligned}$$

There are three of these, and hence the quotient ring $\mathbb{Z}_6/K_2$ has three elements and must be isomorphic to $\mathbb{Z}_3$; examination of the additive and multiplicative structure confirms this.

The next example concerns evaluation homomorphisms in the polynomial ring $\mathbb{Z}[x]$.

Example 9.44 Let $\text{ev}_0: \mathbb{Z}[x] \rightarrow \mathbb{Z}$ be the evaluation homomorphism at $x = 0$. Then $K = \ker(\text{ev}_0)$ consists of all polynomials $p \in \mathbb{Z}[x]$ for which $\text{ev}_0(p) = p(0) = 0$. These are the polynomials with zero constant term; that is, polynomials of the form $p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x$. This is exactly the principal ideal $\langle x \rangle$ generated by x .

The cosets of K are of the form $q(x) + K$, but after a little thought it becomes apparent that they are determined entirely by their constant terms and are thus all of the form $a_0 + K$ for some $a_0 \in \mathbb{Z}$.

The additive structure of the quotient ring $\mathbb{Z}[x]/K$ is just ordinary integer addition, since $(a_0 + K) + (b_0 + K) = (a_0 + b_0) + K$; furthermore the multiplicative structure is also the same as integer multiplication, since $(a_0 + K) \cdot (b_0 + K) = (a_0 b_0) + K$. Hence $\mathbb{Z}[x]/K \cong \mathbb{Z}$.

This is also an example of the First Isomorphism Theorem for rings. Also, recall from Section 3.2 that the process of passing to a quotient group G/H can be viewed as making everything in the normal subgroup H equal to each other, and to the identity, with whatever contextual meaning that might have. For example, factoring the dihedral group D_3 by the rotation subgroup R_3 effectively glues together those points in the equilateral triangle under investigation that differ by a $\frac{2\pi}{3}$ rotation.

Essentially the same thing happens with quotient rings: we choose an ideal I in a ring R , make all of the elements of I in some sense “the same”, and see what happens to R as a result. In Example 9.44 we took the quotient of $\mathbb{Z}[x]$ by the principal ideal $\langle x \rangle$, which equated x with 0. This killed any occurrences of x or its higher powers, leaving just the constant terms. The next example explores this idea further.

Example 9.45 Let $\mathbb{Z}[x]$ as usual denote the ring of polynomials in x over $\mathbb{Z}$, and let $\langle 2 \rangle$ be the principal ideal generated by 2; this consists of all polynomials with even integer coefficients.

The cosets of $\langle 2 \rangle$ in $\mathbb{Z}[x]$ are all of the form $p + \langle 2 \rangle$, where $p(x) = a_n x^n + \cdots + a_1 x + a_0$, with $a_0, \dots, a_n \in \{0, 1\}$.

Hence the quotient ring $\mathbb{Z}[x]/\langle 2 \rangle$ consists of the (cosets determined by) finite-degree polynomials in x with coefficients all equal to either 0 or 1. By considering the additive and multiplicative structure of these cosets, it turns out that $\mathbb{Z}[x]/\langle 2 \rangle \cong \mathbb{Z}_2[x]$, the ring of polynomials in x over $\mathbb{Z}_2$.

In this example, factoring by the ideal $\langle 2 \rangle$ has the effect of setting 2 equal to 0, but leaving x and its powers unchanged. This results in a change of coefficient ring from $\mathbb{Z}$ to $\mathbb{Z}_2$.

Example 9.46 We can factor the ring $\mathbb{Z}[x]$ by more complicated ideals. In particular, $I = \langle x^2+1 \rangle$ consists of all polynomials over $\mathbb{Z}$ which have (x^2+1) as a factor. In the quotient $\mathbb{Z}[x]/\langle x^2+1 \rangle$, two polynomials are equivalent if they differ by a multiple of x^2+1 .

But also, since $x^2 + 1 \equiv 0$, we have $x^2 \equiv -1$. This means that any $p \in \mathbb{Z}[x]$ maps in the quotient $\mathbb{Z}[x]/\langle x^2+1 \rangle$ to the corresponding polynomial with every x^2 replaced by a -1 . So, for example, $x^6 + 3x^5 + 2x^2 - 1$ maps to $(-1)^3 + 3x(-1)^2 + 2(-1) - 1 = 3x - 4$.

The cosets in $\mathbb{Z}[x]/\langle x^2+1 \rangle$ are thus all of the form $(a+bx)+I$, where $a, b \in \mathbb{Z}$. Their additive structure is much as we'd expect, with

$$((a+bx)+I) + ((c+dx)+I) = ((a+c) + (b+d)x) + I$$

for any $a, b, c, d \in \mathbb{Z}$. The multiplicative structure is interesting and somewhat familiar:

$$\begin{aligned} ((a+bx)+I) \cdot ((c+dx)+I) &= ((a+bx) \cdot (c+dx)) + I \\ &= (ac + (ad+bc)x + bdx^2) + I \\ &= ((ac-bd) + (ad+bc)x) + I. \end{aligned}$$

In fact, the quotient ring $\mathbb{Z}[x]/\langle x^2+1 \rangle$ is isomorphic to the ring $\mathbb{Z}[i]$ of Gaussian integers, via the isomorphism $(a+bx)+I \mapsto a+bi$.

The following example concerns group rings.⁷

⁷ Example 8.31, page 331.

Example 9.47 Let $f: D_3 \rightarrow \mathbb{Z}_2$ be the group homomorphism that maps $e, \rho, \rho^2 \mapsto 0$ and $m_1, m_2, m_3 \mapsto 1$. Then $\ker(f) = R_3$, the rotation subgroup. This group homomorphism f defines an **induced homomorphism** $f_*: \mathbb{Z}D_3 \rightarrow \mathbb{Z}\mathbb{Z}_2$ with kernel $K = \ker(f_*) = \mathbb{Z}R_3$.

The elements of this kernel are formal sums of the form $ae + b\rho + c\rho^2$, where $a, b, c \in \mathbb{Z}$, and the quotient ring $\mathbb{Z}D_3/K$ consists of the cosets

$$\begin{aligned} K = e+K = \rho+K = \rho^2+K &= \{ae + b\rho + c\rho^2 : a, b, c \in \mathbb{Z}\} \\ m_1+K = m_2+K = m_3+K &= \{am_1 + bm_2 + cm_3 : a, b, c \in \mathbb{Z}\}. \end{aligned}$$

The quotient ring thus has two elements and is isomorphic to $\mathbb{Z}_2$.

By Proposition 4.29 we know that all kernels of group homomorphisms are normal, and all normal subgroups are kernels of group homomorphisms. We've already seen that kernels of ring homomorphisms are ideals, and the next proposition restates this fact and proves the converse too.

Proposition 9.48 Let $f: R \rightarrow S$ be a (possibly nonunital) ring homomorphism. Then $\ker(f)$ is an ideal of R . Furthermore, any ideal I of a (possibly nonunital) ring R can be expressed as the kernel of some ring homomorphism $f: R \rightarrow S$.

Proof Let $f: R \rightarrow S$ be a (possibly nonunital) ring homomorphism.

Then $K = \ker(f)$ is an ideal: we know from the discussion following Definition 9.24 that $(K, +)$ is an additive subgroup of $(R, +)$, and that K is closed under multiplication; furthermore Proposition 9.28 confirms the absorption conditions.

To show the converse, as with Proposition 4.29 we use the quotient homomorphism $q: R \rightarrow R/I$ that maps every element $a \in R$ to the coset $a+I$ in R/I . Then $I = \ker(q)$. $\square$

Examples 9.42 and 9.44, as noted earlier, suggest an analogue for rings of the First Isomorphism Theorem for groups,⁸ which we now state and prove.

⁸ Theorem 4.40, page 117.

Theorem 9.49 (First Isomorphism Theorem) *Let $f: R \rightarrow S$ be a (possibly nonunital) ring homomorphism, and let $K = \ker(f)$. Then the function $\phi: R/K \rightarrow \text{im}(f)$ given by $\phi(a+K) = f(a)$ is an isomorphism. That is,*

$$R/\ker(f) \cong \text{im}(f).$$

The following proof is analogous to the group-theoretic case.

Proof As before, we need to show that ϕ is a well-defined function, a ring homomorphism, and a bijection.

Suppose that $a, b \in R$ such that the cosets $a+K$ and $b+K$ coincide. Then $(a-b)$ must lie in the kernel K , and hence $f(a-b)$ lies in the image $f(K) = \{0\}$. Hence $f(a-b) = 0$, and since f is a ring homomorphism,

$$0 = f(a-b) = f(a) - f(b),$$

therefore $f(a) = f(b)$. So if $a+K = b+K$ it follows that

$$\phi(a+K) = f(a) = f(b) = \phi(b+K),$$

and thus ϕ is well-defined.

Next we have to check that ϕ is a ring homomorphism. To do this we have to check both the addition and multiplication conditions. Firstly,

$$\begin{aligned} \phi((a+K) + (b+K)) &= \phi((a+b)+K) = f(a+b) \\ &= f(a) + f(b) = \phi(a+K) + \phi(b+K) \end{aligned}$$

and secondly,

$$\begin{aligned} \phi((a+K) \cdot (b+K)) &= \phi((a \cdot b)+K) = f(a \cdot b) \\ &= f(a) \cdot f(b) = \phi(a+K) \cdot \phi(b+K). \end{aligned}$$

Every element of $\text{im}(f)$ is of the form $\phi(a+K)$ for some coset $(a+K) \in R/K$, so ϕ is surjective.

To see that ϕ is injective, suppose that $\phi(a+K) = \phi(b+K)$ for some $a, b \in R$. Then

$$0 = \phi(a+K) - \phi(b+K) = f(a) - f(b) = f(a-b)$$

and hence $(a - b) \in K = \ker(f)$, which means that $a + K = b + K$, and therefore ϕ is injective.

Thus $\phi: R/K \rightarrow \text{im}(f)$ is a well-defined ring isomorphism. $\square$

As with the group-theoretic version, this is an important and useful theorem. In Example 9.46 we saw that $\mathbb{Z}[x]/\langle x^2+1 \rangle$ was isomorphic to the ring $\mathbb{Z}[i]$ of Gaussian integers.⁹ We can use the First Isomorphism Theorem to prove this result too.

⁹ Example 8.6, page 320.

Example 9.50 The evaluation homomorphism $\text{ev}_i: \mathbb{Z}[x] \rightarrow \mathbb{Z}[i]$ maps a polynomial $p \in \mathbb{Z}[x]$ to its evaluation $p(i) \in \mathbb{Z}[i]$. This is surjective: we can find a polynomial $a+bx \in \mathbb{Z}[x]$ that evaluates to any given Gaussian integer $a+bi \in \mathbb{Z}[i]$. The kernel $\ker(\text{ev}_i)$ consists of all polynomials that have (x^2+1) as a factor; this is the principal ideal $\langle x^2+1 \rangle$. Hence, the First Isomorphism Theorem tells us that

$$\mathbb{Z}[x]/\langle x^2+1 \rangle = \mathbb{Z}[x]/\ker(\text{ev}_i) \cong \text{im}(\text{ev}_i) = \mathbb{Z}[i].$$

There are ring-theoretic versions of the Second¹⁰ and Third Isomorphism Theorems,¹¹ which we now state and prove.

¹⁰ Theorem 4.62, page 126.

¹¹ Theorem 4.65, page 128.

First we need some preliminary facts about ideals and subrings. Here and in what follows, for some subring S and ideal I in a ring R , we set

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

by analogy with Definition 2.10.

Proposition 9.51 *Let R be a (unital or nonunital) ring, let S be a (possibly nonunital) subring of R , and let I be an ideal of R . Then*

- (i) $S+I$ is a (possibly nonunital) subring of R ,
- (ii) $S \cap I$ is an ideal of S , and
- (iii) I is an ideal of $S+I$.

Proof First we check that $S+I$ is a subring of R . To check condition SR1, that $S+I$ is an additive subgroup of R , we observe that $S+I$ contains $0 = 0 + 0$ since both S and I do. Also, for any $s \in S$ and $i \in I$ the element $s+i$ has additive inverse $-(s+i) = (-s)+(-i) \in S+I$. Furthermore, $S+I$ is closed under addition because both S and I are. To check SR2, closure under multiplication, observe that for any $a, b \in S$ and $i, j \in I$, we have

$$(a+i) \cdot (b+j) = (a+i) \cdot b + (a+i) \cdot j = a \cdot b + (i \cdot b + a \cdot j + i \cdot j).$$

This is an element of $S+I$ since $a \cdot b \in S$ and all of $i \cdot b, a \cdot j, i \cdot j \in I$.

If S has a unity element 1 then $S+I$ has unity element $1+0$, since

$$(1+0) \cdot (a+i) = 1 \cdot a + 0 \cdot a + 1 \cdot i + 0 \cdot i = a + i,$$

$$(a+i) \cdot (1+0) = a \cdot 1 + i \cdot 1 + a \cdot 0 + i \cdot 0 = a + i.$$

Hence $S+I$ is a (possibly nonunital) subring of R .

To see that $S \cap I$ is an ideal of S , we first check that it is an additive subgroup of S . Both S and I contain 0 so their intersection does. If $a \in S \cap I$ then $-a \in S$ and $-a \in I$, so $-a \in S \cap I$ as well. And $S \cap I$ is closed under addition: if $a, b \in S \cap I$ then $a+b \in S$ and $a+b \in I$, since both S and I are closed under addition, and therefore $a+b \in S \cap I$.

Finally, we need to check the absorption conditions for $S \cap I$. If $a \in S$ and $s \in S \cap I$, then both $a \cdot s$ and $s \cdot a$ lie in S since S is closed under multiplication, and both $a \cdot s$ and $s \cdot a$ also lie in I since I is an ideal of R . Hence $a \cdot s \in S \cap I$, and therefore $S \cap I$ is an ideal of S .

The third statement, that I is an ideal of $S+I$, is also just a matter of checking definitions. Since I is an ideal of R it is already an additive subgroup of R , and since it is a subset of $S+I$ it must therefore also be an additive subgroup of $S+I$. All that remains is to check the left and right absorption conditions. For any $s \in S$ and $i, j \in I$ we have

$$(a+i) \cdot j = a \cdot j + i \cdot j \in I, \quad j \cdot (a+i) = j \cdot a + j \cdot i \in I$$

and hence I is an ideal of $S+I$. □

We're now ready to prove the Second Isomorphism Theorem for rings. As with the group-theoretic version of the Second Isomorphism Theorem,¹² we prove this using the First Isomorphism Theorem.¹³

Theorem 9.52 (Second Isomorphism Theorem) *Let R be a ring and let I be an ideal of R . Then*

$$(R+I)/I \cong R/(R \cap I).$$

Proof If we can construct a surjective ring homomorphism $f: S \rightarrow (S+I)/I$ with kernel $S \cap I$ then we can simply apply the First Isomorphism Theorem to get the desired result.

Recall that $S+I$ consists of elements of R which admit a (not necessarily unique) decomposition as a sum $s+i$ for some elements $s \in S$ and $i \in I$. Then by Proposition 9.51, I is an ideal of $S+I$ and we can thus form the quotient $(S+I)/I$, which consists of cosets of the form $(s+i)+I$. But since $i \in I$, each such coset is equal to $s+I$. So, we want a surjective ring homomorphism that maps elements of S to cosets of the form $s+I$ in the quotient $(S+I)/I$.

An obvious candidate is the function $f: S \rightarrow (S+I)/I$ mapping an element $s \in S$ to the corresponding coset $s+I$.

This is a ring homomorphism, since for any $a, b \in S$,

$$f(a+b) = (a+b)+I = (a+I) + (b+I) = f(a) + f(b)$$

and

$$f(a \cdot b) = (a \cdot b)+I = (a+I) \cdot (b+I) = f(a) \cdot f(b).$$

¹² Theorem 4.62, page 126.

¹³ Theorem 9.49, page 364.

It is also surjective, since for any coset $s+I$ in $(S+I)/I$ there is an element $s \in S$ with $f(s) = s+I$.

Finally, for $s \in S$ we have $f(s) = 0+I$ exactly when $s+I = 0+I$, which is true if and only if $s \in I$. Hence $s \in S \cap I$ and therefore $\ker(f) = S \cap I$.

Thus

$$S/S \cap I = S/\ker(f) \cong \text{im}(f) = (S+I)/I$$

by the First Isomorphism Theorem. $\square$

Example 9.53 Let $R = \mathbb{Z}[i]$, the ring of Gaussian integers, let $I = \langle 3i \rangle$, and let $S = \mathbb{Z}$.

Then I is the principal ideal generated by $3i$, consisting of all elements of the form $3i \cdot z$ for $z \in \mathbb{Z}[i]$. This is equal to the subset $3\mathbb{Z}[i] = \{3a + 3bi : a, b \in \mathbb{Z}\}$ of Gaussian integers with real and imaginary parts a multiple of 3.

Now $S+I$ is the set of elements of the form $n+z$, where $n \in \mathbb{Z}$ and $z \in I$. This consists of all Gaussian integers with imaginary part a multiple of 3, so $S+I = \{a + 3bi : a, b \in \mathbb{Z}\}$. It is a subring of $R = \mathbb{Z}[i]$ by Proposition 9.51(i). The cosets of I in $S+I$ are therefore I , $1+I$ and $2+I$, so $(S+I)/I \cong \mathbb{Z}_3$.

The intersection $S \cap I$ comprises the elements of I that have zero imaginary part, namely $\{\dots, -6, -3, 0, 3, 6, \dots\} = 3\mathbb{Z}$. The quotient $S/(S \cap I)$ is thus $\mathbb{Z}/3\mathbb{Z} \cong \mathbb{Z}_3$.

So, in this case we have $(S+I)/I \cong \mathbb{Z}_3 \cong S/(S \cap I)$ as predicted by the Second Isomorphism Theorem.

There is also a Third Isomorphism Theorem for rings, which again we prove using the corresponding First Isomorphism Theorem.

Theorem 9.54 (Third Isomorphism Theorem) *Let R be a ring, and let I and J be ideals of R such that $I \subseteq J$. Then*

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

Proof Let $f: R/J \rightarrow R/I$ such that $f(a+J) = a+I$ for any $a \in R$. This map is well-defined, since if $a+J = b+J$ it follows that $a-b \in J \subseteq I$, so $a+I = b+I$. It is also a ring homomorphism since

$$\begin{aligned} f((a+J) + (b+J)) &= f((a+b)+J) = (a+b)+I \\ &= (a+I) + (b+I) = f(a+J) + f(b+J) \end{aligned}$$

and

$$\begin{aligned} f((a+J) \cdot (b+J)) &= f((a \cdot b)+J) = (a \cdot b)+I \\ &= (a+I) \cdot (b+I) = f(a+J) \cdot f(b+J). \end{aligned}$$

Now the kernel $\ker(f) = \{a+J : a+I = I\} = \{a+J : a \in I\}$, which is just the quotient I/J . Furthermore, the image $\text{im}(f) = \{a+I : a \in R\}$

is simply the quotient R/I . Then

$$(R/J)/(I/J) = (R/J)/\ker(f) \cong \operatorname{im}(f) = R/I$$

by the First Isomorphism Theorem. $\square$

Example 9.55 Let $R = \mathbb{Z}[i]$, $I = \langle 3i \rangle$ and $J = \langle 9i \rangle$. Then $R/I \cong \mathbb{Z}_3[i]$ and $R/J \cong \mathbb{Z}_9[i]$. Furthermore, $I/J \cong \mathbb{Z}_3[i]$, and we have

$$(R/J)/(I/J) \cong \mathbb{Z}_9[i]/\mathbb{Z}_3[i] \cong \mathbb{Z}_3[i] \cong R/I$$

as predicted by the Third Isomorphism Theorem.

The prime ideal is a princess of the world of ideals. Her father is the prince “Point” in the world of geometry. Her mother is the princess “Prime Numbers” in the world of numbers. She inherits the purity from her parents.

— Kazuya Kato,
lecture, 11 February 2013

9.4 Prime ideals and maximal ideals

WE SAW EARLIER that $n\mathbb{Z} = \langle n \rangle$ is an ideal of $\mathbb{Z}$. Indeed, all ideals of $\mathbb{Z}$ are of this form. Also, we have a series of inclusions $\langle 9 \rangle \subset \langle 3 \rangle \subset \mathbb{Z}$. Is there another ideal that fits between $\langle 3 \rangle$ and $\mathbb{Z}$? That is, can we find an ideal I such that $\langle 3 \rangle \subset I \subset \mathbb{Z}$?

The answer is no. Suppose that I strictly contains $\langle 3 \rangle$, and therefore that there exists some nonzero integer $a \in I \setminus \langle 3 \rangle$. This means that a is not of the form $3n$ for any integer n . Suppose that $a = 3n+1$ for some $n \in \mathbb{Z}$. Then we need I to satisfy the conditions for an additive abelian subgroup of $\mathbb{Z}$, and hence I must also contain all elements of the form $3k+1$ for $k \in \mathbb{Z}$, because it has to be closed under addition. It must contain all elements of the form $3k+2$ as well, since the additive inverse of $3n+1$ is $-(3n+1) = -3n-1 = -3(n+1)+2$. Thus $I = \mathbb{Z}$. If, on the other hand, a is of the form $3n+2$, then $I = \mathbb{Z}$ by a very similar argument. We give ideals of this type a special name:

Definition 9.56 Let R be a (possibly nonunital) ring, and let I be a proper ideal of R . Then I is a **maximal ideal** if there exists no other ideal J such that $I \subset J \subset R$.

Let $a \in R$ be some element of R . Then the principal ideal $\langle a \rangle$ is a **maximal principal ideal** if there exists no other principal ideal $\langle b \rangle$ such that $\langle a \rangle \subset \langle b \rangle \subset R$.

Here are some examples of maximal and non-maximal ideals. First let's look at a matrix ring.

Example 9.57 The ideal $\langle \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \rangle$ in $M_2(\mathbb{Z})$ consists of all matrices of the form $\begin{bmatrix} 2a & 2b \\ 3c & 3d \end{bmatrix}$, where $a, b, c, d \in \mathbb{Z}$. It is contained in $\langle \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \rangle$ and is hence not a maximal ideal of $M_2(\mathbb{Z})$.

The next example considers the ideals of a finite ring, $\mathbb{Z}_{12}$.

Example 9.58 The ring $\mathbb{Z}_{12}$ has six ideals, namely

$$\begin{aligned}\langle 0 \rangle &= \{0\}, & \langle 6 \rangle &= \{0, 6\}, & \langle 4 \rangle &= \{0, 4, 8\}, \\ \langle 3 \rangle &= \{0, 3, 6, 9\}, & \langle 2 \rangle &= \{0, 2, 4, 6, 8, 10\}, & \langle 1 \rangle &= \mathbb{Z}_{12}.\end{aligned}$$

The lattice diagram in Figure 9.2 highlights the fact that only $\langle 2 \rangle$ and $\langle 3 \rangle$ are maximal ideals of $\mathbb{Z}_{12}$.

It's interesting to ask what the corresponding quotient rings are, and it turns out that

$$\begin{aligned}\mathbb{Z}_{12}/\langle 0 \rangle &\cong \mathbb{Z}_{12}, & \mathbb{Z}_{12}/\langle 6 \rangle &\cong \mathbb{Z}_6, \\ \mathbb{Z}_{12}/\langle 4 \rangle &\cong \mathbb{Z}_4, & \mathbb{Z}_{12}/\langle 3 \rangle &\cong \mathbb{Z}_3, \\ \mathbb{Z}_{12}/\langle 2 \rangle &\cong \mathbb{Z}_2, & \mathbb{Z}_{12}/\langle 1 \rangle &\cong 0.\end{aligned}$$

Of these, only $\mathbb{Z}_2$ and $\mathbb{Z}_3$ are fields, and these are the ones corresponding to maximal ideals.

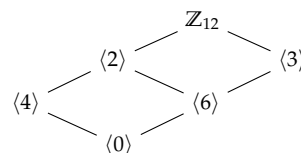


Figure 9.2: Lattice diagram for the ring $\mathbb{Z}_{12}$

All of the ideals of $\mathbb{Z}_{12}$ are principal. In the next chapter we will study a class of rings with this property, called principal ideal domains.

Example 9.59 The ideal $\langle t^2 \rangle$ in $\mathbb{Z}[t]$ consists of all polynomials with zero constant and linear terms; that is, those of the form

$$a_n t^n + \cdots + a_2 t^2.$$

This is not a maximal ideal, because it is properly contained in $\langle t \rangle$, which is also not a maximal ideal. However, the ideal

$$\begin{aligned}\langle t, 2 \rangle &= \{a_n t^n + \cdots + a_1 t : a_1, \dots, a_n \in \mathbb{Z}\} \\ &\cup \{b_n t^n + \cdots + b_1 t + b_0 : b_0, \dots, b_n \in 2\mathbb{Z}\}\end{aligned}$$

is maximal, as is any other ideal of the form $\langle t, p \rangle$ for p prime. The quotients are

$$\mathbb{Z}[t]/\langle t \rangle \cong \mathbb{Z}, \quad \mathbb{Z}[t]/\langle t, 2 \rangle \cong \mathbb{Z}_2.$$

Of these, $\mathbb{Z}_2$ is a field, while $\mathbb{Z}$ is just an integral domain. The quotient $\mathbb{Z}[t]/\langle t^2 \rangle$ is not even that, since $t^2 = 0$ means that t is a zero divisor.

However, if we consider the ring $\mathbb{R}[t]$ of polynomials with real coefficients, we find that the ideal $\langle x \rangle$ is maximal: it consists of the polynomials with zero constant term, and no other proper ideal of $\mathbb{R}[t]$ contains it. The quotient $\mathbb{R}[t]/\langle t \rangle \cong \mathbb{R}$, which is a field.

In fact, not only is $\langle t^2 \rangle$ not maximal in $\mathbb{Z}[t]$, no ideal of the form $\langle f \rangle$ for $f \in \mathbb{Z}[t]$ is either:

Proposition 9.60 *No principal ideal in $\mathbb{Z}[t]$ is maximal.*

Proof Let $n \in \mathbb{Z}_* \subset \mathbb{Z}[t]$ be a nonzero integer. If $n = \pm 1$ then $\langle n \rangle = \mathbb{Z}[t]$ isn't maximal. If $n \neq \pm 1$ then $\langle n \rangle \subset \langle n, t \rangle$ and hence again

$\langle n \rangle$ isn't maximal.

Now suppose that $f = a_m t^m + \cdots + a_1 t + a_0 \in \mathbb{Z}[t]$ is a polynomial of degree $m > 0$. Then $\langle f \rangle$ isn't maximal either, since the ideal $\langle f, p \rangle$ strictly contains it, where p is prime and $p \nmid a_m$. $\square$

This correlation between maximal ideals and fields holds in general:

Proposition 9.61 *Let R be a commutative ring, and I be an ideal of R . Then R/I is a field if and only if I is maximal.*

Proof Suppose I is maximal in R , and let $a \in R \setminus I$. Then $(a+I)$ isn't the additive identity $(0+I) = I$ in the quotient R/I . We want to show that $(a+I)$ is a unit in R/I , and to do this we need to find some $b \in R \setminus I$ such that

$$(a+I) \cdot (b+I) = (1+I) = (b+I) \cdot (a+I).$$

Let $J = \{ra + i : r \in R, i \in I\}$. Then J is an additive abelian group, since

$$(r_1 a + i_1) + (r_2 a + i_2) = (r_1 + r_2)a + (i_1 + i_2) \in J,$$

the additive identity $0 = 0a + 0 \in J$, and for any $(ra + i) \in J$ we have

$$-(ra + i) = (-r)a + (-i) \in J.$$

Furthermore, J satisfies the left and right absorption conditions, since for any $s \in R$,

$$s \cdot (ra + i) = (s \cdot r)a + s \cdot i \in J$$

$$\text{and} \quad (ra + i) \cdot s = ra \cdot s + i \cdot s = (r \cdot s)a + i \cdot s \in J$$

Finally, J is closed under multiplication, since

$$\begin{aligned} (r_1 a + i_1) \cdot (r_2 a + i_2) &= (r_1 r_2 a) \cdot a + i_1 r_2 a + r_1 a i_2 + i_1 i_2 \\ &= (r_1 r_2 a)a + (i_1 r_2 a + r_1 a i_2 + i_1 i_2) \in J. \end{aligned}$$

Hence J is an ideal of R . But J contains I , since every element of $i \in I$ can be written in the form $0a + i$. But I is maximal in R , and hence $J = I$.

In particular, $1 \in J = I$, so there exists some $b \in R$ and $i \in I$ such that $1 = ba + i$. Hence $ba + i \in (1+I)$, so $(b+I) = (a+I)^{-1}$ as required, and therefore R/I is a field.

Conversely, suppose that R/I is a field, and J is an ideal of R that strictly contains I . Let $a \in J \setminus I$. Then $(a+I)$ is a nonzero element of R/I . Since R/I is a field, it follows that $(a+I)$ is invertible, and hence there exists a coset $(b+I)$ such that $(a+I) \cdot (b+I) = (1+I)$. Since $a \in J$ and J is an ideal, it follows that $ab \in J$. Hence $(a+I)(b+I) = 1+I$ and so $ab+I = 1+I$, which means that $(1-ab) \in I \subset J$, so $1 = (1-ab)+ab \in J$.

Hence $J = R$ by Proposition 9.38, and so I is maximal. $\square$

We can now characterise the maximal ideals of $\mathbb{Z}$. Recall that all additive subgroups of $\mathbb{Z}$ are of the form $n\mathbb{Z}$ for $n \in \mathbb{N}$. These are exactly the principal ideals $\langle n \rangle$, and the quotient $\mathbb{Z}/\langle n \rangle \cong \mathbb{Z}_n$ is a field exactly when n is prime. Therefore, by Proposition 9.61, the maximal ideals of $\mathbb{Z}$ are exactly the principal ideals $\langle p \rangle$ where p is prime.

This leads us in an interesting direction. Suppose that $r, s \in \mathbb{Z}$ and their product $rs \in \langle p \rangle = p\mathbb{Z}$. Then if p is prime, p must divide either r or s .

So, the maximal ideals of $\mathbb{Z}$ are exactly those principal ideals $\langle n \rangle$ satisfying this property: that if $r, s \in \mathbb{Z}$ and $rs \in \langle n \rangle$, then either r or s must be a multiple of n , or equivalently r or s must lie in $\langle n \rangle$.

Is this true in general? Suppose we have a commutative ring R with an ideal I such that for any $r, s \in R$ with $r \cdot s \in I$, either $r \in I$ or $s \in I$. Does it necessarily follow that I is maximal?

The answer is no, as the following example shows.

Example 9.62 Let $R = \mathbb{Z}[x]$. Then $\langle x \rangle$ is not maximal, since $\langle x \rangle \subset \langle x, 2 \rangle$. But for any $p, q \in \mathbb{Z}[x]$ such that $pq \in \langle x \rangle$ it must be the case that either $p \in \langle x \rangle$ or $q \in \langle x \rangle$.

The reason for this is that $\langle x \rangle$ consists of polynomials with zero constant term. Suppose that $p = a_mx^m + \cdots + a_1x + a_0$ and $q = b_nx^n + \cdots + b_1x + b_0$. Then

$$pq = a_mb_nx^{m+n} + \cdots + (a_1b_0 + a_0b_1)x + a_0b_0.$$

This product lies in $\langle x \rangle$ if and only if $a_0b_0 = 0$. Since $\mathbb{Z}$ is an integral domain, this can only be the case if either $a_0 = 0$ or $b_0 = 0$, in which case either $p \in \langle x \rangle$ or $q \in \langle x \rangle$.

This seems to be an interesting concept, worth investigating further, and to that end we introduce the following definition.

Definition 9.63 Let R be a commutative ring, and let I be an ideal of R . We say I is **prime** if, for any $a, b \in R$ with $ab \in I$, it must also be the case that either $a \in I$ or $b \in I$ (or both).

So, the above example demonstrates the existence of ideals that are prime but not maximal, and we know from the preceding discussion of $\mathbb{Z}$ that some prime ideals are maximal too. The obvious next question is to ask whether all maximal ideals are prime, but the answer to this requires a little more work.

By Proposition 9.61 we know that for any commutative ring R and ideal I , the quotient R/I is a field exactly when I is maximal. In the above example, the quotient $\mathbb{Z}[x]/\langle x \rangle$ is isomorphic to $\mathbb{Z}$, which isn't a field. But in some sense it's next best thing: an integral domain. It

transpires that this is true in general: quotients of commutative rings by prime ideals are integral domains.

Proposition 9.64 *Let R be a commutative ring, and let I be an ideal of R . Then R/I is an integral domain if and only if I is prime.*

Proof Suppose R/I is an integral domain, and that $a, b \in R$ such that $a \cdot b \in I$. Then $(a+I)(b+I) = (ab+I) = I$, which is the additive identity in R/I . Since R/I is an integral domain, this means that either $(a+I) = I$, in which case $a \in I$, or $(b+I) = I$, in which case $b \in I$. Hence I is prime.

Conversely, suppose that I is prime. We must show that R/I has no zero divisors. For nonzero $a, b \in R$ such that $(a+I)(b+I) = (0+I) = I$, we have $(ab+I) = I$, and hence $ab \in I$. Then, since I is prime, either $a \in I$, in which case $(a+I) = I$ is the zero element in R/I , or $b \in I$, in which case $(b+I) = I$ is the zero element in R/I . Thus R/I is an integral domain. $\square$

This is all we need to answer the question of whether maximal ideals are prime: they are.

Corollary 9.65 *Maximal ideals are prime.*

Proof Let I be a maximal ideal in some commutative ring R . Then R/I is a field, and hence also an integral domain by Proposition 8.47. Therefore I is a prime ideal by Proposition 9.64. $\square$

Summary

IN CHAPTERS 2 AND 3 we investigated subgroups and quotient groups, in the process learning some useful techniques for understanding the internal structure of groups. In this chapter, we sought to develop analogous concepts and techniques to help us to better understand rings.

We began by introducing the idea of a **subring**,¹⁴ a subset $S \subseteq R$ that forms a ring in its own right via the addition and multiplication operations inherited from the ambient ring R . Almost immediately we ran into some notational technicalities, related to the lack of a universal convention on whether rings necessarily have unity elements. We adopted the convention that by default a **subring** has a unity element, a **nonunital subring** doesn't, and a **unital subring** not only has a unity element, but it must have the *same* unity element as the ambient ring. (This is not universal: some books adopt a different convention.) For S to be a subring of R , we require that $(S, +)$ be a subgroup of

¹⁴ Definition 9.2, page 345.

$(R, +)$, that S be closed under multiplication, and that S have a unity element.¹⁵

¹⁵ Proposition 9.3, page 345.

For example, the ring $\mathbb{Z}_4$ has no proper *unital* subrings, although the subset $\{0, 2\}$ forms a subring in the sense just described, with 2 serving as a unity element.¹⁶ This is a special case of the more general fact that neither $\mathbb{Z}$ nor $\mathbb{Z}_n$ have any proper unital subrings,¹⁷ although $\mathbb{Z}$ has countably many nonunital subrings, all of the form $n\mathbb{Z}$.¹⁸ However, any unital ring R contains a subring isomorphic to either $\mathbb{Z}$ or $\mathbb{Z}_n$ for some integer $n > 1$,¹⁹ this is called the **prime subring**, and is the smallest nontrivial unital subring of R . It is also the intersection of all unital subrings of R ; indeed any intersection of unital subrings is also a unital subring.²⁰

¹⁶ Example 9.4, page 346.

¹⁷ Proposition 9.5, page 346.

¹⁸ Proposition 9.6, page 347.

¹⁹ Proposition 9.7, page 347.

²⁰ Proposition 9.11, page 350.

In general, the direct product $G \times H$ of two groups G and H will have at least one subgroup isomorphic to G and one isomorphic to H .²¹ The corresponding statement is not always true for rings, because for a direct sum $R \oplus S$, the obvious inclusion functions $i_1: R \hookrightarrow R \oplus S$ and $i_2: S \hookrightarrow R \oplus S$ don't necessarily map unity elements to unity elements. More precisely, any injective ring homomorphism $f: R \rightarrow R \oplus S$ must map 0 to $(0, 0)$ and 1 to $(1, 1)$, which the canonical inclusion functions don't do.

²¹ Example 4.16, page 106.

However, a useful result about direct products of groups that does work in a ring-theoretic context is the fact that $\mathbb{Z}_m \oplus \mathbb{Z}_n \cong \mathbb{Z}_{mn}$ if and only if $\gcd(m, n) = 1$,²² this is a version of the celebrated Chinese Remainder Theorem.²³ A corollary of this is that the ring $\mathbb{Z}_n$ has a unique (up to permutation of factors) decomposition as a direct sum of rings of the form $\mathbb{Z}_{p^k}$ where p is prime.²⁴

²² Proposition 1.32, page 16.

²³ Proposition 9.8, page 348.

²⁴ Corollary 9.9, page 349.

A matrix group $M_n(R)$ typically has numerous subrings: the subrings $UT_n(R)$, $LT_n(R)$ and $D_n(R)$ of triangular and diagonal matrices,²⁵ the prime subring $\langle I \rangle$ generated by the identity matrix I , and also any ring of the form $M_n(S)$ where S is a subring of R .

²⁵ Example 9.10, page 349.

In order to define and investigate quotient rings, we spent some time devising the correct ring-theoretic analogue of a normal subgroup. Bearing in mind that normal subgroups are equivalent to kernels of group homomorphisms,²⁶ we began by studying some basic properties of ring homomorphisms.

²⁶ Proposition 4.29, page 113.

For a homomorphism $f: R \rightarrow S$, where R is unital, it isn't necessarily the case that the unity element in R maps to the unity element in S ; indeed, S need not be unital. It must, however, be the case that $f(1)^2 = f(1)$; that is, ring homomorphisms map unity elements to **idempotent** elements.²⁷ This led us to introduce the notions of a **unital homomorphism**, that maps unity to unity, and a **nonunital homomorphism**, that either maps unity to some other idempotent

²⁷ Proposition 9.16, page 351.

²⁸ Definition 9.17, page 352.

element, or whose domain or codomain are nonunital rings.²⁸ Any nontrivial ring R admits at least one unital homomorphism, the identity homomorphism $\text{id}_R: R \rightarrow R$ that maps everything to itself, and one nonunital homomorphism, the zero homomorphism $z_R: R \rightarrow R$ that maps everything to 0.

²⁹ Example 9.20, page 353.

An interesting class of homomorphisms are the **evaluation homomorphisms**, defined on polynomial rings.²⁹ For some fixed element $a \in R$, the homomorphism $\text{ev}_a: R[x] \rightarrow R$ maps a polynomial $p \in R[x]$ to its evaluation $p(a)$ obtained by setting $x = a$.

³⁰ Proposition 9.21, page 353.

Next we examined the properties of images and kernels of ring homomorphisms. Images of group homomorphisms are always subrings of their codomains, but the situation is slightly more complicated with ring homomorphisms. In general, images of nonunital homomorphisms are (possibly nonunital) subrings, while images of unital homomorphisms are unital subrings.³⁰ The **kernel** of a ring homomorphism $f: R \rightarrow S$ is the set of elements of R that are mapped to the zero element in S .³¹ Kernels of ring homomorphisms are at least nonunital subrings of their domains: they are additive abelian subgroups and closed under multiplication, but need not have a unity element. They do, however, satisfy **absorption conditions**;³² that is, for any $r \in R$ and $k \in \ker(f)$, the products $r \cdot k$ and $k \cdot r$ lie in $\ker(f)$.

³¹ Definition 9.24, page 355.

³² Proposition 9.28, page 356.

Motivated by this, we defined an **ideal** I of a ring R to be a subset of R that satisfies the first two subring conditions and the left and right absorption conditions.³³ Ideals are the ring-theoretic analogues of normal subgroups. In particular, for a commutative unital ring R and some element $a \in R$ we defined the **principal ideal** $\langle a \rangle$ generated by a to be the set $\{ra : r \in R\}$.³⁴ Just as every group has at least two normal subgroups, itself and the trivial group, every ring has at least two ideals: itself and $\{0\}$.^{35,36} If R is a unital ring, then the only ideal of R that contains 1 is R itself;³⁷ more generally, R is the only ideal containing a unit.³⁸ That is, no proper ideals contain any invertible elements or the unity element 1. As anticipated, not only is every ring homomorphism kernel an ideal, every ideal can be regarded as the kernel of some ring homomorphism.³⁹

³³ Definition 9.29, page 356.

³⁴ Example 9.30, page 357.

³⁵ Example 9.34, page 358.

³⁶ Example 9.35, page 358.

³⁷ Proposition 9.38, page 358.

³⁸ Proposition 9.39, page 359.

³⁹ Proposition 9.48, page 363.

Having obtained a definition for the ring-theoretic analogue of a normal subgroup, we then turned our attention to cosets. There were two approaches we could have taken, the additive or multiplicative viewpoint, but for various reasons the multiplicative one was unsuitable for our purposes. So, given a ring R , an ideal I and an element $a \in R$, we defined the **coset** of I corresponding to a to be the set $a + I = \{a + i : i \in I\}$.⁴⁰ This enabled us to define the **quotient ring** R/I ,⁴¹ whose elements are the cosets of I in R , with addition and

⁴⁰ Definition 9.40, page 359.

⁴¹ Proposition 9.41, page 360.

multiplication operation defined by

$$(a+I) + (b+I) = (a+b)+I,$$

$$(a+I) \cdot (b+I) = (a \cdot b)+I.$$

Ring-theoretic versions of the First, Second and Third Isomorphism Theorems hold.^{42,43,44}

⁴² Theorem 9.49, page 364.

Next we considered a couple of special types of ideal. A **maximal ideal** I in a ring R is a proper ideal of R that is contained in no larger proper ideal: any other ideal J strictly containing I must be the full ring R itself.⁴⁵ There is a similar notion of a **maximal principal ideal**:

⁴³ Theorem 9.52, page 366.

⁴⁴ Theorem 9.54, page 367.

a *principal* ideal contained in no larger proper *principal* ideal. Note that a maximal principal ideal need not be a maximal ideal, and in particular no principal ideal in $\mathbb{Z}[t]$ is maximal.⁴⁶ A proper ideal

⁴⁵ Definition 9.56, page 368.

$I \subset R$ is maximal if and only if the quotient R/I is a field.⁴⁷ A related concept is that of a **prime ideal**: an ideal $I \subset R$ is **prime** if for any

⁴⁶ Proposition 9.60, page 369.

⁴⁷ Proposition 9.61, page 370.

$a, b \in R$ with $ab \in I$, then either $a \in I$ or $b \in I$.⁴⁸ An ideal I is prime if and only if the quotient R/I is an integral domain.⁴⁹ Hence, since all fields are integral domains, all maximal ideals are prime.⁵⁰

⁴⁸ Definition 9.63, page 371.

⁴⁹ Proposition 9.64, page 372.

⁵⁰ Corollary 9.65, page 372.

References and further reading

Exercises

- 9.1 Show that the quotient ring $\mathbb{Z}[x, y] / \langle x^2+1, y^2+1, xy+yx \rangle$ is isomorphic to the ring $\mathbb{L}$ of Lipschitz integers described in Example 8.12.
- 9.2 Suppose that $1 < k < n$ and that $\gcd(k, n) > 1$. Show that $\langle k \rangle$ is a nonunital subring of $\mathbb{Z}_n$, and that all nonunital subrings of $\mathbb{Z}_n$ are of this form.
- 9.3 Let I denote the 3×3 identity matrix, and let A be a matrix of the form

$$A = \begin{bmatrix} 0 & a & b \\ 0 & 0 & c \\ 0 & 0 & 0 \end{bmatrix}$$

for some nonzero $a, b, c \in R$. Then $A^3 = 0$ but neither A nor A^2 are zero. We say that A is **nilpotent**; it has finite multiplicative order. Show that I and A generate a commutative subring of $M_3(R)$.

- 9.4 Show that the evaluation homomorphism $\text{ev}_a: R[x] \rightarrow R$ is indeed a ring homomorphism for any $a \in R$.
- 9.5 Show that, for a commutative unital ring R and element $a \in R$, the principal ideal $\langle a \rangle$ is indeed an ideal of R .

10 Domains

IN THIS CHAPTER WE WILL STUDY some important types of integral domains. The ring $\mathbb{Z}$ of integers was our original example of an integral domain, and it has some special properties that we will generalise to other contexts.

In particular, the Division Theorem¹ holds in $\mathbb{Z}$: given any integers a and b , we can find integers q and r , with $0 \leq r < b$ such that $a = qb + r$. We want to extend this idea to other rings and find out in exactly which circumstances it holds.

In the last chapter we devised and studied the concept of an **ideal**, a (possibly nonunital) subring satisfying **absorption conditions**.² In particular we met a class of ideals called **principal ideals**,³ of a commutative ring R which are of the form $\langle a \rangle = \{ra : r \in R\}$ for some $a \in R$. We will see later that every ideal of $\mathbb{Z}$ is principal; that is, of the form $\langle n \rangle$ for some $n \in \mathbb{Z}$. We want to know whether this is significant, and if so whether there are any other rings that have only principal ideals.

Finally, every integer has a unique (up to permutation and multiplication by ± 1) factorisation into prime integers. We will generalise the notion of a prime element to other rings, and investigate what sort of rings have unique prime factorisations for arbitrary elements.

10.1 Euclidean domains

THE RING $\mathbb{Z}$ OF INTEGERS satisfies the Division Theorem; this useful fact has been known for at least 2400 years, and probably longer. We now want to generalise it to other rings, and to formulate some criteria that will tell us exactly which rings have an analogous property. This is important, because the Division Theorem essentially says we can do the kind of long division with remainders that we learned about in primary school. Sometimes we will want to do this in other rings,

There are several requirements that must be met to establish a domain. In general it must be responsibly managed. There must be a responsible person to serve as a coordinator for domain related questions, there must be a robust name service, it must be of at least a minimum size, and the domain must be registered with the central domain administrator.

— Jon Postel (1943–1998),
The Domain Names Plan and Schedule,
IETF RFC 881, November 1983

¹ Theorem A.37, page 499.

² Definition 9.29, page 356.

³ Example 9.30, page 357.

Besides, they relate that Euclid was asked by Ptolemy whether there was any shorter way to the attainment of geometry than by his elementary institution, and that he answered there was no other royal path which led to geometry.

— Proclus Lycæus (422–485AD),
A Commentary on the First Book of
Euclid's Elements,
translated by Thomas Taylor
(1758–1835) (1788)

such as the rings $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ of Gaussian and Eisenstein integers, or the ring $\mathbb{Q}[x]$ of polynomials with rational coefficients, and we need to know if this is going to be possible.

But straight away we run into a problem. Given two integers $a, b \in \mathbb{Z}$ with $b \neq 0$, the Division Theorem asserts the existence of two other integers $q, r \in \mathbb{Z}$ such that $a = qb + r$ and $0 \leq r < b$. The first part of this is fine: in any ring it makes sense to talk about some element a being equal to an expression $qb + r$, because all we need is addition and multiplication, both of which are defined in any ring. The difficulty arises with the second part: the requirement that $0 \leq r < b$. Not all rings, and certainly not all of the ones we're most interested in, have a canonical ordering $<$. For example, it's not immediately obvious what it might mean to say that $3x^2 + 2 < 4x^3 + x - 1$ in $\mathbb{Q}[x]$, or that $(2 - i) < (5 + 3i)$ in $\mathbb{Z}[i]$.

Things become slightly clearer if we restate the Division Theorem for $\mathbb{Z}$ in a slightly different way:

Theorem 10.1 (Division Theorem for $\mathbb{Z}$) *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. Then there exist integers $q, r \in \mathbb{Z}$ such that $a = qb + r$ and either $r = 0$ or $|r| < |b|$.*

We can do almost all of this in an arbitrary ring R ; again the complication is the last bit, the need to extend or provide a suitable alternative to the absolute value function $|\cdot|$ so that the whole statement makes sense in our chosen ring. Ultimately, what we want from an absolute value is that it maps every element of the ring to a non-negative real number. However, our proof of the Division Theorem for $\mathbb{Z}$ relies on induction, so really we need our generalised absolute value function to map into $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ rather than $\mathbb{R}_{\geq 0}$, at least if we're to avoid some complicated technical digressions.

So, we need a function that maps from our ring R into $\mathbb{N}_0$. What else do we need it to do? Well, while studying linear algebra, functional analysis or metric spaces you may have met a similar function called a **norm**. Typically, we require such a function to satisfy the **triangle inequality**

$$|\mathbf{u} + \mathbf{v}| \leq |\mathbf{u}| + |\mathbf{v}|, \quad (10.1)$$

but in this case we're not going to: it turns out that this is overly strict for our purposes. Instead, we require

$$|ab| \geq |a| \quad (10.2)$$

for any $a, b \in R$. The reason for this will become apparent soon.

What we want, then, is a sort of generalised absolute value function, or norm, that maps into $\mathbb{N}_0$, satisfies (10.2), and for which the corresponding Division Theorem holds.

If our ring R has one of these functions that enables us to state and prove an analogue of the Division Theorem, then that means we can happily do long division with remainders in R . But the statement of the Division Theorem really depends on our choice of norm function, and there may be more than one of these. (Indeed, the function $n \mapsto n^2$ suffices in $\mathbb{Z}$ as well as the usual $|\cdot|$.) So it makes more sense to fold the Division Theorem condition into the criteria for our norm function, and thus we arrive at the following definition:

Definition 10.2 Let R be an integral domain. A **Euclidean norm** or **Euclidean valuation** is a function $v: R_* \rightarrow \mathbb{N}_0$ such that

$$v(ab) \geq v(a) \quad (10.3)$$

for all $a, b \in R_*$, and for any $a, b \in R$ with $b \neq 0$ there exist $q, r \in R$ such that $a = qb + r$ and either $r = 0$ or $v(r) < v(b)$.

The term **Euclidean norm** is also often used to refer to the Euclidean distance function in $\mathbb{R}^n$ determined by Pythagoras' Theorem, and so from now on we will use the term **Euclidean valuation** to avoid confusion.

The integral domains we're interested in at the moment are those for which we can find a Euclidean valuation:

Definition 10.3 An integral domain R that admits a Euclidean valuation is said to be a **Euclidean domain**.

Note that for a given integral domain R to be a Euclidean domain, we only need to be able to define a Euclidean valuation $v: R_* \rightarrow \mathbb{N}_0$: any suitable one will do, and the valuation v isn't itself considered part of the structure for the domain. That is, the mere existence of a Euclidean valuation is enough.

So, the ring $\mathbb{Z}$ of integers is a Euclidean domain via the absolute value function $v(n) = |n|$, and also via $v(n) = n^2$.

The significance of the condition (10.3) is given by the next proposition:

Proposition 10.4 Let R be a Euclidean domain, and let $v: R_* \rightarrow \mathbb{N}_0$ be a Euclidean valuation for R . Then $v(1) \leq v(a)$ for any $a \in R_*$, and $v(a) = v(1)$ if and only if a is a unit.

Proof The condition (10.3) tells us that, in particular,

$$v(a) = v(1a) \geq v(1)$$

for any $a \in R_*$.

If $a \in R$ is a unit, then

$$v(1) = v(aa^{-1}) \geq v(a),$$

but we've just seen that $v(a) \geq v(1)$, so $v(a) = v(1)$.

Now suppose that $a \in R_*$ such that $v(a) = v(1)$. Then by the Division Theorem condition, there exists $q, r \in R$ such that

$$1 = qa + r$$

and either $r = 0$ or $v(r) < v(a) = v(1)$. But we already know that $v(r)$ can't be strictly less than $v(1)$, so it must be the case that $r = 0$, and hence $1 = qa$, which means that $q = a^{-1}$, and therefore a is a unit. $\square$

What other examples of Euclidean domains can we find? Well, probably the next objects that we learn how to long divide are polynomials. As remarked earlier, there isn't a canonical ordering on the ring $\mathbb{F}[x]$ for a given field $\mathbb{F}$, nor is there an obvious absolute value function we can use. However, any nonzero polynomial $p(x) \in \mathbb{F}[x]$ does have a specific non-negative integer associated with it: its **degree** $\deg(p)$.

This yields a function $\deg: \mathbb{F}[x]_* = \mathbb{F}[x] \setminus \{0\} \rightarrow \mathbb{N}_0$ that certainly satisfies the condition (10.3), since for any $p, q \in \mathbb{F}[x]$ with $\deg(p) = m \geq 0$ and $\deg(q) = n \geq 0$, we have $\deg(pq) = m + n \geq m$.

The corresponding Division Theorem takes a little work to prove, but follows a similar pattern to the one for $\mathbb{Z}$:

Theorem 10.5 *Let $\mathbb{F}$ be a field, and let $f, g \in \mathbb{F}[x]$ with $g \neq 0$. Then there exist unique polynomials $q, r \in \mathbb{F}[x]$ such that $f = qg + r$, where either $r = 0$ or $\deg(r) < \deg(g)$.*

Proof First we prove the existence of q and r . If $f = 0$ or $\deg(f) < \deg(g)$ then we set $q = 0$ and $r = f$. Otherwise, we proceed by induction on $\deg(f)$.

If $\deg(f) = \deg(g) = 0$ then both f and g are nonzero constants, and hence they are elements of $\mathbb{F}$, and therefore units. We can thus set $q = fg^{-1}$ and $r = 0$.

Suppose that

$$f = a_mx^m + \cdots + a_1x + a_0 \quad \text{and} \quad g = b_nx^n + \cdots + b_1x + b_0,$$

so that $\deg(f) = m$ and $\deg(g) = n$. Suppose also that $m > 0$ and that $0 \leq n \leq m$. Furthermore, we assume that the theorem holds for all f and g such that $\deg(f) < m$.

Now we apply long division to get a polynomial

$$\begin{aligned} h &= f - a_mb_n^{-1}x^{m-n}g \\ &= (a_mx^m + \cdots + a_1x + a_0) \\ &\quad - (a_mx^m + a_mb_n^{-1}x^{m-1} + \cdots + a_mb_n^{-1}b_0x^{m-n}) \\ &= (a_{m-1} - a_mb_n^{-1}b_{n-1})x^{m-1} + \cdots \end{aligned}$$

and we see that either $\deg(h) < m$ or $h = 0$.

If $h = 0$ then set $q = a_m b_n^{-1} x^{m-n}$ and $r = 0$ to get $f = qg$.

Otherwise, if $h \neq 0$, then $\deg(h) < m$ and so by induction we can find $s, t \in \mathbb{F}[x]$ where $h = sg + t$, with $\deg(t) < \deg(g)$. Then

$$\begin{aligned} f &= a_m b_n^{-1} x^{m-n} g + sg + t \\ &= (a_m b_n^{-1} x^{m-n} + s)g + t \end{aligned}$$

and we can thus set $q = a_m b_n^{-1} x^{m-n} + s$ and $r = t$.

To prove uniqueness, suppose that $f = q_1 g + r_1 = q_2 g + r_2$ with either $r_1 = 0$ or $\deg(r_1) < \deg(g)$, and either $r_2 = 0$ or $\deg(r_2) < \deg(g)$. Then

$$(q_1 g + r_1) - (q_2 g + r_2) = 0,$$

which means that

$$(r_1 - r_2) = (q_2 - q_1)g.$$

Either $r_1 - r_2 = 0$, in which case $r_1 = r_2$ and $q_1 = q_2$, or

$$\deg(r_1 - r_2) = \deg(q_2 - q_1) \deg(g)$$

which means that $\deg(r_1 - r_2) \geq \deg(g)$, which can't happen.

Hence $r_1 = r_2$ and $q_1 = q_2$ as claimed. $\square$

So $\mathbb{F}[x]$ is a Euclidean domain via the map $\deg: \mathbb{F}[x]_* \rightarrow \mathbb{N}_0$. This means we can, in principle at least, do long division with polynomials.

Example 10.6 Let $f = x^4 + 5x^3 - 2x - 3$ and $g = x^2 + 1$ in $\mathbb{Q}[x]$, the ring of univariate polynomials with rational coefficients. We want to find $q, r \in \mathbb{Q}[x]$ such that $f = qg + r$. Long division proceeds as follows:

$$\begin{array}{r} x^2 + 5x - 1 \\ x^2 + 1 \overline{) x^4 + 5x^3 + 0x^2 - 2x - 3} \\ \underline{x^4 + 0x^3 + x^2} \\ 5x^3 - x^2 - 2x \\ \underline{5x^3 + 0x^2 + 5x} \\ -x^2 - 7x - 3 \\ \underline{-x^2 + 0x - 1} \\ -7x - 2 \end{array}$$

The quotient is the polynomial left on the top line, namely $q = x^2 + 5x - 1$, while the remainder is the polynomial left at the end, in this case $r = -7x - 2$. We can easily check this:

$$\begin{aligned} qg + r &= (x^2 + 5x - 1)(x^2 + 1) + (-7x - 2) \\ &= x^4 + 5x^3 + 5x - 1 - 7x - 2 \\ &= x^4 + 5x^3 - 2x - 3 \\ &= f \end{aligned}$$

We can use this to prove an important fact about polynomial rings over fields.

Theorem 10.7 (Remainder Theorem) *Let $\mathbb{F}$ be a field, let $f \in \mathbb{F}[x]$ be a polynomial over $\mathbb{F}$, and let $a \in \mathbb{F}$. Then $f(a) = 0$ if and only if $(x-a)$ divides f with no remainder.*

Proof By Theorem 10.5 we know that $\mathbb{F}[x]$ is a Euclidean domain. Therefore

$$f = (x-a)q + r$$

where $q, r \in \mathbb{F}[x]$ and either $r = 0$ or $\deg(r) < \deg(x-a) = 1$. Hence r is a constant polynomial; that is, an element $r \in \mathbb{F}$. Substituting $x = a$ into the above Euclidean decomposition, we get

$$f(a) = (a-a)q(a) + r(a) = 0 + r = r.$$

So $f(a) = 0$ exactly when the remainder $r = 0$. □

An important corollary of this theorem gives an upper bound on the number of roots a polynomial $f \in \mathbb{F}[x]$ can have.

Corollary 10.8 *Let $\mathbb{F}$ be a field, and let $f \in \mathbb{F}[x]$. Then there are at most $\deg(f)$ distinct elements $a \in \mathbb{F}$ for which $f(a) = 0$.*

Proof We proceed by induction on the degree of f . If $\deg(f) = 0$ then $f \neq 0$ is a nonzero constant in $\mathbb{F}$, and so $f(a) \neq 0$ for any $a \in \mathbb{F}$. Now suppose that $\deg(f) = k$ and that any polynomial of degree less than k has at most $(k-1)$ distinct roots in $\mathbb{F}$. If $f(a) = 0$ then by Theorem 10.7 we can factorise $f = (x-a)q$ where $\deg(q) = k-1$, and by the inductive hypothesis, q has at most $(k-1)$ roots in $\mathbb{F}$. Thus f has at most $k = \deg(f)$ roots in $\mathbb{F}$. □

A useful application of this concerns the group of units of a field.

Proposition 10.9 *Let $\mathbb{F}$ be a field, and let $U(\mathbb{F}) = \mathbb{F} \setminus \{0\}$ denote its (multiplicative) group of units. Any finite subgroup of $U(\mathbb{F})$ is cyclic.*

Proof Let $G \leq U(\mathbb{F})$ be some finite subgroup of $U(\mathbb{F})$ with order $|G| = n$, and suppose it isn't cyclic. Then G must be Abelian, and by Corollary 5.49 we know it must be isomorphic to a direct sum $\mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_k}$ of finite cyclic groups, where $m_1 | m_2 | \cdots | m_k$. Let $m = m_k$ and consider some arbitrary element $(g_1, \dots, g_k) \in \mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_k}$. We then have

$$(g_1, \dots, g_k)^m = (g_1^m, \dots, g_k^m) = (1, \dots, 1),$$

and so $g^m = 1$ for any $g \in G$.

But this means that any of the n elements of G is a solution of the equation $g^m = 1$, and hence a root of the polynomial $f = x^m - 1$ in $\mathbb{F}[x]$. However, $n = m_1 \times \cdots \times m_k > m_k = m = \deg(f)$, which contradicts Corollary 10.8.

Therefore G must be cyclic. $\square$

The next corollary is a special case, which we will use in Section 10.A to help characterise the prime elements in the ring $\mathbb{Z}[i]$ of Gaussian integers.

Corollary 10.10 *Let $\mathbb{F}_p$ denote the finite field of prime order p . Then $U(\mathbb{F}_p) \cong \mathbb{Z}_{p-1}$.*

Proof By Lemma 2.43, the group $U(\mathbb{F}_p) = \mathbb{Z}_p^\times$ consists of those integers k such that $0 < k < p$ and $\gcd(k, p) = 1$. The number of such integers is given by Euler's totient function ϕ ,⁴ and Proposition 2.46 tells us that $\phi(p) = p-1$. Hence $|U(\mathbb{F}_p)| = p-1$, and Proposition 10.9 says that it must be cyclic, so $U(\mathbb{F}_p) = \mathbb{Z}_p^\times \cong \mathbb{Z}_{p-1}$ as claimed. $\square$

⁴ Definition 2.44, page 63.

The ring $\mathbb{Z}[i]$ of Gaussian integers is also a Euclidean domain. To show this, we need to find a suitable valuation function $v: \mathbb{Z}[i] \rightarrow \mathbb{N}_0$, and prove the corresponding Division Theorem.

Our motivating example of a valuation was the absolute value function $|\cdot|: \mathbb{Z} \rightarrow \mathbb{N}_0$, which we can view as the restriction to $\mathbb{Z} \subset \mathbb{R}$ of the absolute value function $|\cdot|: \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ for the real numbers. The Gaussian integers can be viewed as a subset of the complex number field $\mathbb{C}$, so we might be inclined to use the restriction to $\mathbb{Z}[i]$ of the complex modulus function $|\cdot|: \mathbb{C} \rightarrow \mathbb{R}_{\geq 0}$. Unfortunately, this doesn't quite work, because the modulus of an arbitrary Gaussian integer need not be a nonnegative integer: for example, $|1+i| = \sqrt{2}$. In some sense, this should be fine, because the set of positive algebraic irrational numbers are countable, and hence there exists a bijection to $\mathbb{N}_0$, which we could compose with the modulus to get the desired function, but it'll be much simpler if we can find an alternative valuation that maps directly into $\mathbb{N}_0$. Happily, we can: $v: \mathbb{Z}[i] \rightarrow \mathbb{N}_0$ given by $v(a+bi) = |a+bi|^2 = a^2 + b^2$.

Proposition 10.11 *The ring $\mathbb{Z}[i]$ of Gaussian integers is a Euclidean domain via the valuation $v(a+bi) = a^2 + b^2$.*

Proof The Gaussian integers form an integral domain: they're commutative and none of them are zero divisors. So we need to show that the map v is a Euclidean valuation.

Suppose that $z = a+bi$ and $w = c+di$. Then

$$\begin{aligned} v(zw) &= v((a+bi)(c+di)) = v((ac-bd) + (ad+bc)i) \\ &= (ac-bd)^2 + (ad+bc)^2 = a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2 \\ &= (a^2+b^2)(c^2+d^2) \geq a^2+b^2 = v(z) \end{aligned}$$

To prove the Division Theorem is, as usual, slightly more involved. As is often the case, it's helpful to think of the problem geometrically.

The Gaussian integers are exactly those points in the complex plane with integer coordinates. The valuation v is then the square of the modulus; it gives the distance from the origin to the point representing the Gaussian integer in question.

The principal ideal $\langle w \rangle = \langle c + di \rangle$ can be regarded as a lattice of points in $\mathbb{Z}[i] \subset \mathbb{C}$. Converting to polar coordinate, there exists some $p \in \mathbb{R}_{\geq 0}$ and $\theta \in [0, 2\pi)$ such that $w = pe^{i\theta}$. The lattice corresponding to $\langle w \rangle$ can then be constructed by taking the full lattice $\mathbb{Z}[i]$, scaling it by a factor of p , and rotating it through an angle θ .

Now, for any other Gaussian integer $z = a + bi$, there is some point in this lattice $\langle w \rangle$ that is closest to it. More precisely, there exists some Gaussian integer $u \in \langle w \rangle$ such that $v(z - u) \leq \frac{1}{2}p^2$. But if u is an element of $\langle w \rangle$, then that means there exists some other Gaussian integer $q \in \mathbb{Z}[i]$ such that $u = qw$. So qw is within squared distance $\frac{1}{2}p^2$ of z ; that is, $v(qw - z) \leq \frac{1}{2}p^2$. To find q , consider

$$s = \frac{z}{w} = \frac{a + bi}{c + di} = \frac{(a + bi)(c - di)}{c^2 + d^2} = \frac{z\bar{w}}{v(w)}.$$

This will, in general, not be in $\mathbb{Z}[i]$, but it will at most be distance $\frac{1}{\sqrt{2}}$ from something that is. Pick the nearest such Gaussian integer and call it q , so that $v(s - q) \leq \frac{1}{2}$.

Hence there exists some other Gaussian integer $r = z - qw$ such that either $r = 0$ (if s does happen to lie in $\mathbb{Z}[i]$), or

$$v(r) = v(z - qw) \leq \frac{1}{2}p^2 < p^2 = v(w).$$

Thus v is a Euclidean valuation on $\mathbb{Z}[i]$, and therefore $\mathbb{Z}[i]$ is a Euclidean domain. $\square$

This is best illustrated by an example.

Example 10.12 Suppose we want to find the quotient q and remainder r of $14+2i$ when divided by $5+3i$. To find q we want the closest Gaussian integer to

$$\frac{14+2i}{5+3i} = \frac{(14+2i)(5-3i)}{25+9} = \frac{76-32i}{34} = \frac{38}{17} - \frac{16}{17}i \approx 2.24 + 0.94i.$$

The closest Gaussian integer to this is $2-i$, so we set $q = 2-i$. Then

$$r = (14+2i) - (2-i)(5+3i) = (14+2i) - (13+i) = 1+i.$$

Hence $14+2i = (2-i)(5+3i) + (1+i)$, and also

$$v(r) = v(1+i) = 2 < 34 = v(5+3i) = v(w)$$

as required. This is depicted geometrically in Figure 10.1.

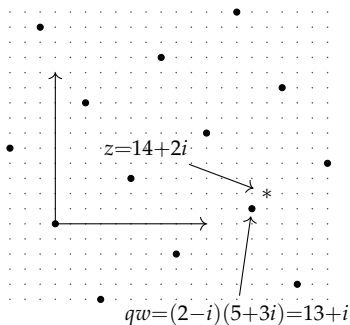


Figure 10.1: The Division Theorem in $\mathbb{Z}[i]$. The small dots represent the Gaussian integers, while the larger dots represent the ideal $\langle w \rangle = \langle 5+3i \rangle$.

The Division Theorem in $\mathbb{Z}$ is the main ingredient for Euclid's Algorithm for finding the greatest common divisor of two integers. There's nothing stopping us from doing the same in any other Euclidean

domain, but we're then faced with the following question: what is the significance of the end result of this process? Ideally, we'd hope that it's a suitable analogue of a greatest common divisor in whatever Euclidean domain we happen to be working in.

10.2 Divisors, primes and irreducible elements

IN ORDER TO EXPLORE THIS IDEA, we first need to decide what we mean by a **divisor** in some arbitrary integral domain. This is fairly straightforward, and we arrive quickly at the following definition.

Definition 10.13 Let R be an integral domain, and let $a, b \in R$. Then a **divides** b , or is a **divisor** or **factor** of b if there exists some element $r \in R$ such that $b = ra$. We denote this by $a|b$.

This can be extended to noncommutative division rings by the introduction of **left** and **right divisors**. We say that a is a **left divisor** of b if there exists some $r \in R$ such that $b = ar$, and a **right divisor** of b if there exists some $s \in R$ for which $b = sa$. Furthermore, a is a **(two-sided) divisor** of b if it is both a left and right divisor.

We will concentrate on the commutative case, in integral domains, from now on. In $\mathbb{Z}$, for example, 2 divides 6 since there exists some integer, namely 3, such that $6 = 2 \times 3$. Furthermore, every multiple of 6 is even, and hence a multiple of 2. In the last chapter we met an elegant way of thinking of this in terms of ideals: the even numbers and the multiples of 6 both form principal ideals of $\mathbb{Z}$. We'll investigate principal ideals more in the next section, but for now we can rephrase the observation that 6 is even as $6 \in \langle 2 \rangle$, and the fact that every multiple of 6 is even as $\langle 6 \rangle \subseteq \langle 2 \rangle$. This is true in general:

Proposition 10.14 Let R be an integral domain. Then the following statements are equivalent, for any $a, b \in R$:

- (i) $a|b$
- (ii) $b \in \langle a \rangle$
- (iii) $\langle b \rangle \subseteq \langle a \rangle$

Proof Suppose that $a|b$. Then there exists some $r \in R$ such that $b = ra$. But the principal ideal $\langle a \rangle$ consists of all elements of R of the form ra for some $r \in R$. Hence $b \in \langle a \rangle$.

Now suppose $b \in \langle a \rangle$. Then $b = ra$ for some $r \in R$, and any other multiple of b (that is, any element of the form sb for some $s \in R$) can be written as $sb = s(ra) = (sr)a \in \langle a \rangle$. But $\langle b \rangle$ consists of all elements of the form sb , so any element of $\langle b \rangle$ must also lie in $\langle a \rangle$, so $\langle b \rangle \subseteq \langle a \rangle$.

Prime numbers are those which have no other factor but one, as three only has a third, and five only has a fifth, and seven only has a seventh; that is, they have only one factor. Composite numbers are those which not only divide by one, but are also produced by another number, such as nine, twenty-one, fifteen and twenty-five. That is, we say three times three, and seven times three, three times five, and five times five.

— St Isidore of Seville (c.560–636AD),
Etymologiae III:7

Finally, suppose $\langle b \rangle \subseteq \langle a \rangle$. This means that every element of the form sb is also of the form ra for some $r, s \in R$. In particular, setting $s = 1$, we find $b = 1b = ra$ for some $r \in R$. Therefore $b = ra$ and thus $a|b$. $\square$

Is it possible for two elements of an integral domain to divide each other? In the case of $\mathbb{Z}$ this only happens in very specific circumstances: for two integers $a, b \in \mathbb{Z}$ we have $a|b$ and $b|a$ only when $a = \pm b$. But for two polynomials $f, g \in \mathbb{Q}[x]$ we see this more often: $f|g$ and $g|f$ only when $f = qg$ for some nonzero $q \in \mathbb{Q}$.

What do these two examples have in common? The answer is that two elements of the given domains divide each other exactly when one is a multiple of the other by an invertible element: $\mathbb{Z}$ has two elements ± 1 , while the units in $\mathbb{Q}[x]$ are the nonzero constant polynomials, which are exactly the nonzero rational numbers $\mathbb{Q}_*$.

Definition 10.15 Let R be an integral domain. Then two elements $a, b \in R$ are **associates** if $a|b$ and $b|a$. We denote this by $a \sim b$.

Proposition 10.16 Let R be an integral domain. Then the following statements are equivalent for any $a, b \in R$:

- (i) $a \sim b$,
- (ii) $\langle a \rangle = \langle b \rangle$, and
- (iii) $a = qb$ for some unit $q \in R$.

Proof The equivalence of (i) and (ii) follows from Proposition 10.14.

If $a \sim b$ then $a|b$, which means that $\langle b \rangle \subseteq \langle a \rangle$. Also $b|a$, which means that $\langle a \rangle \subseteq \langle b \rangle$. Hence $\langle a \rangle = \langle b \rangle$. Each step in this argument is reversible, so the converse holds too: if $\langle a \rangle = \langle b \rangle$ then $a \sim b$.

Now suppose $a \sim b$. If $a = 0$ then $b = 0$, so suppose both a and b are nonzero. Then there exist $r, s \in R$ such that $a = rb$ and $b = sa$. Then $a = r(sa)$, so $(rs)a - a = 0$ and hence $(rs - 1)a = 0$. Since R is an integral domain, it has no zero divisors, and therefore $rs = 1$. Thus r and s are units, with $s = r^{-1}$. The converse is straightforward: if there exists a unit $q \in R$ such that $a = qb$ then $b|a$, and also $b = q^{-1}a$ so $a|b$. Thus $a \sim b$. $\square$

Proposition 10.17 Association is an equivalence relation.

Proof Association is reflexive: $a \sim a$ for any $a \in R$, since $a = 1a$.

Association is symmetric: if $a \sim b$ then $a = ub$ for some unit $u \in R$. Then $b = u^{-1}a$, and hence $b \sim a$.

Association is transitive: if $a \sim b$ and $b \sim c$ then $a = ub$ and $b = vc$ for some units $u, v \in R$. Then $a = (uv)c$, and thus $a \sim c$, since uv is also a unit of R . $\square$

We want to extend the familiar notions of **greatest common divisors**

and **least common multiples** to arbitrary integral domains, in particular those which might not have an obvious concept of “greatest” or “least”. To start with, we revisit the usual definitions in $\mathbb{Z}$:

Definition 10.18 For any two integers $a, b \in \mathbb{Z}$, an integer $d \in \mathbb{Z}$ is a **common divisor** of a and b if $d|a$ and $d|b$. It is a **greatest common divisor** if for any other common divisor $e \in \mathbb{Z}$, we have $0 < |e| \leq |d|$. Furthermore, an integer $m \in \mathbb{Z}$ is a **common multiple** of a and b if $a|m$ and $b|m$. It is a **least common multiple** if for any other common multiple $n \in \mathbb{Z}$ we have $|n| \geq |m|$.

This certainly makes sense in $\mathbb{Z}$ and we can see how to extend it to any other Euclidean domain:

Definition 10.19 Let R be a Euclidean domain, let $a, b \in R$ and let v be a Euclidean valuation for R . Then $d \in R$ is a **common divisor** for a and b if $d|a$ and $d|b$. It is a **greatest common divisor** if for any other common divisor $e \in R$, we have $0 < v(e) \leq v(d)$. Furthermore, $m \in R$ is a **common multiple** of a and b if $a|m$ and $b|m$. It is a **least common multiple** if for any other common multiple $n \in R$ we have $v(n) \geq v(m)$.

This definition is fine for Euclidean domains, but we want one that works in an arbitrary integral domain. (However, having a definition that makes sense doesn’t guarantee that such elements will actually exist: we will later see examples of integral domains where arbitrary elements don’t always have a greatest common divisor.)

The key observation is that if $a, b \in \mathbb{Z}$ and $d \in \mathbb{Z}$ is a greatest common divisor for a and b , then not only does any other common divisor $e \in \mathbb{Z}$ satisfy $0 < |e| < |d|$, but it must also be the case that $e|d$ as well.

This leads to the following more general definition.

Definition 10.20 Let R be an integral domain and let $a, b \in R$. Then $d \in R$ is a **common divisor** for a and b if $d|a$ and $d|b$. It is a **greatest common divisor** if for any other common divisor $e \in R$ we have $e|d$. An element $m \in R$ is a **common multiple** of a and b if $a|m$ and $b|m$, and a **least common multiple** if for any other common multiple $n \in R$ it is the case that $m|n$.

More generally, for any finite collection $a_1, \dots, a_n$ of elements of R , we say that $d \in R$ is a **greatest common divisor** of $a_1, \dots, a_n$ if $d|a_1, \dots, d|a_n$ and for any other common divisor $e \in R$ with $e|a_1, \dots, e|a_n$ we have $e|d$. Similarly, l is a **least common multiple** of $a_1, \dots, a_n$ if $a_1|l, \dots, a_n|l$ and for any other common multiple $m \in R$ with $a_1|m, \dots, a_n|m$ we have $l|m$.

Where the meaning is clear from context, we will often just refer to “the” greatest common divisor, which we will then denote $\gcd(a, b)$ or $\gcd(a_1, \dots, a_n)$, and “the” least common multiple, which we will then write as $\text{lcm}(a, b)$ or $\text{lcm}(a_1, \dots, a_n)$.

Some books refer to the **highest common factor** instead of the greatest common divisor, which is then denoted $\text{hcf}(a, b)$ or $\text{hcf}(a_1, \dots, a_n)$. Also, some books denote the greatest common divisor by (a, b) or $(a_1, \dots, a_n)$ although this notation is prone to ambiguity and so we won’t use it here.

Definition 10.20 implies Definition 10.19 in a Euclidean domain. Sup-

pose that R is a Euclidean domain and that d is a greatest common divisor for two arbitrary elements $a, b \in R$. Then any other common divisor $e \in R$ must satisfy $e|d$, which means that there exists some $c \in R$ with $d = ce$. Then for any Euclidean valuation v we have $v(d) = v(ce) \geq v(e)$ by (10.3). Similarly, suppose that $m \in R$ is a least common divisor for a and b , and that $n \in R$ is some other common divisor. Then by Definition 10.20 there exists some $l \in R$ such that $n = lm$, and by (10.3) we have $v(n) = v(lm) \geq v(m)$ as required.

We've also been careful to use the indefinite rather than the definite article when discussing greatest common divisors and least common multiples, and the reason for this is that they are not guaranteed to be unique in an arbitrary integral domain. In fact, they're not unique even in $\mathbb{Z}$. For example, both 3 and -3 are greatest common divisors of 6 and 9, and both 18 and -18 are lowest common multiples. What the elements in these pairs have in common is that they differ by a factor of -1 , which is a unit in $\mathbb{Z}$. More generally, they are associates of each other. This is true in any integral domain: any two greatest common divisors are associates, and so are any two least common multiples. This is a consequence of Definitions 10.15 and 10.20.

Any integer $n \in \mathbb{Z}$ is either prime or can be expressed as a finite product of prime factors $n = p_1 \dots p_k$. Later in this chapter we will study factorisation in more detail, but for now it would be useful to have a general notion of primality that works in any integral domain.

The integer 3 is prime. By this, we usually mean that it has no factors except 1 or itself; the integer 4, on the other hand, is **composite** because it factorises as 2×2 . That is, 4 can be expressed as a product of integers which or neither 1 nor 4 itself.

But $3 = (-1) \times (-3)$ as well, so we have to take account of negative integers as well. We can do this by saying that an integer n is prime if any factorisation $n = ab$ requires either a or b to be a unit, and composite otherwise (if it factorises as a product of non-units). Also, we don't usually consider 1 or -1 to be prime or composite, so we'll explicitly discount the case where n is a unit.

There's another way of looking at this, not in terms of what integers divide n , but in terms of what composite integers n itself divides. For example, $3|18$ and 18 can be decomposed as a product in various ways: as 1×18 , as 2×9 or as 3×6 . In each of these cases, 3 divides at least one factor: $3|18$ in the first case, $3|9$ in the second, and $3|3$ and $3|6$ in the last one.

If we try this with a composite integer, by considering how 4 divides 20, then something different happens. In particular, $20 = 1 \times 20 = 2 \times 10 = 4 \times 5$. In one of these factorisations, namely 2×10 , our chosen

composite integer 4 divides neither factor 2 or 10. This can happen because 4 itself factorises into a product 2×2 of non-units.

So this gives a different way of characterising an integer n as prime: it isn't a unit, and for any product $m = ab$ with $n|m$, either $n|a$ or $n|b$.

These definitions are equivalent for integers, but not necessarily in arbitrary integral domains. So we'll use different terms for each:

Definition 10.21 Let R be an integral domain. An element $r \in R$ is **irreducible** if it is not a unit, and if $r = ab$ for some $a, b \in R$ it must be the case that either a or b is a unit.

Slightly confusingly, this is the usual definition of what it means for an integer to be prime, whereas we reserve that term for the following.

Definition 10.22 Let R be an integral domain. An element $r \in R$ is **prime** if it is not a unit, and if for any $a, b \in R$ it must be the case that $r|ab$ implies that either $r|a$ or $r|b$.

These two definitions give rise to subtly different concepts. The subtlety is exacerbated by the fact that in familiar number systems like $\mathbb{Z}$, prime elements are irreducible and irreducible elements are prime. More generally, only the first of these is true in a given integral domain.

Proposition 10.23 Any prime element of an integral domain R is irreducible.

Proof Let $r \in R$ be prime, and suppose that $r = ab$. Clearly $r|r$, so $r|ab$, and since r is prime we must have either $r|a$ or $r|b$. Suppose that $r|a$, without loss of generality. But $a|r$ as well since $r = ab$, so $r \sim a$, which means that $r = ua$ for some unit $u \in R$. By the cancellation law for integral domains (Proposition 8.48) we have $u = b$, so b is a unit and hence r is irreducible. $\square$

In a little while, we'll investigate the exact circumstances in which the converse fails to be true, but for the moment here is an example.

Example 10.24 Let $\mathbb{Z}[\sqrt{-3}] = \{a + b\sqrt{-3} : a, b \in \mathbb{Z}\}$. This is an integral domain, and we claim that in it, 2 is irreducible but not prime. To show this, we need to use the **norm** function $N: \mathbb{Z}[\sqrt{-3}]^* \rightarrow \mathbb{N}$ given by $N(a + b\sqrt{-3}) = a^2 + 3b^2$.

In particular, this function has the following properties:

- (i) $N(r) = 1$ if and only if r is a unit; and
- (ii) $N(rs) = N(r)N(s)$ for any nonzero $r, s \in \mathbb{Z}[\sqrt{-3}]$.

Now observe that $(1 + \sqrt{-3})(1 - \sqrt{-3}) = 4 = 2 \times 2$. Suppose that $2 = rs$ where $r = u_1 + v_1\sqrt{-3}$ and $s = u_2 + v_2\sqrt{-3}$. Then

$$4 = N(2) = N(rs) = N(r)N(s) = (u_1^2 + 3v_1^2)(u_2^2 + 3v_2^2).$$

If $N(r) = 1$ then r is a unit and $r^{-1} = u_1 - v_1\sqrt{-3}$.

Similarly, if $N(s) = 1$ then s is a unit, with $s^{-1} = u_2 - v_2\sqrt{-3}$.

The only other option is that $N(r) = N(s) = 2 \neq 1$. This ensures that neither r nor s are units, but also requires $u_1^2 + 3v_1^2 = 2$, and there are no integers u_1 and v_1 satisfying this equation, so $N(r)$ can't be 2. Thus, the only factorisations $2 = rs$ in $\mathbb{Z}[\sqrt{-3}]$ require either r or s to be a unit. Also, $N(2) = 4 \neq 1$, so 2 itself isn't a unit, and is therefore irreducible.

Now recall that $2|4$ and $4 = (1+\sqrt{-3})(1-\sqrt{-3})$. We know that 2 isn't a unit, so for it to be prime we would require either $2|(1+\sqrt{-3})$ or $2|(1-\sqrt{-3})$. For this to be the case, we would need an element $t \in \mathbb{Z}[\sqrt{-3}]$ for which $2t = 1 \pm \sqrt{-3}$. But this requires $t = \frac{1}{2} \pm \frac{\sqrt{-3}}{2}$, neither of which are elements of $\mathbb{Z}[\sqrt{-3}]$. So 2 is not prime in $\mathbb{Z}[\sqrt{-3}]$.

Something else we notice in this example is that 4 factorises in two different ways $2 \times 2 = (1+\sqrt{-3})(1-\sqrt{-3})$ as a product of irreducible elements in $\mathbb{Z}[\sqrt{-3}]$. In $\mathbb{Z}$, however, we're used to irreducible factorisations being unique (at least up to permutation of factors and multiplication by units $u = \pm 1$). Later we'll return to this idea and investigate when factorisations are unique in this way.

Recall that we defined an ideal $I \subseteq R$ of a ring R to be **prime** if, for any $a, b \in R$ with $ab \in I$, either $a \in I$ or $b \in I$.⁵ This looks very similar to the definition of a prime element in an integral domain R : $p \in R$ is **prime** if, whenever $p|ab$ for some $a, b \in R$, either $p|a$ or $p|b$.⁶

In one definition we are concerned with whether the given ideal I contains one or other element a or b , and in the other we are interested in whether the given element p is a multiple of a or b .

Is this just a coincidental analogy, or is there a deeper connection? And what can we say about irreducible elements and maximal ideals?

Proposition 10.25 *Let p be a nonzero element of an integral domain R . Then*

- (i) *p is irreducible if and only if $\langle p \rangle$ is a maximal principal ideal, and*
- (ii) *p is prime if and only if $\langle p \rangle$ is a prime ideal.*

Proof (i) Suppose that p is irreducible, and $\langle p \rangle$ isn't a maximal principal ideal of R ; that is, there exists some $a \in R$ for which $\langle p \rangle \subsetneq \langle a \rangle \subseteq R$. Then $p \in \langle a \rangle$, so there exists some $b \in R$ for which $p = ab$. Since p is irreducible, either a or b must be a unit. If a is a unit, then $\langle a \rangle = R$ by Proposition 9.39. And if b is a unit, then $a = pb^{-1}$, which must lie in $\langle p \rangle$, so $\langle a \rangle = \langle p \rangle$. In either case, there is no principal ideal between $\langle p \rangle$ and R , so $\langle p \rangle$ is a maximal principal ideal.⁷

⁵ Definition 9.63, page 371.

⁶ Definition 10.22, page 389.

⁷ There may be a non-principal ideal between $\langle p \rangle$ and R , however.

Now suppose $\langle p \rangle$ is a maximal principal ideal of R , and that p is not irreducible. Then $p = ab$ for some $a, b \in R$, neither of which are units. If $a \in \langle p \rangle$ then $a = pc$ for some $c \in R$, and thus $p = ab = pcb$. The cancellation law for integral domains (Proposition 8.48) tells us that $1 = cb$, which means b is a unit, contradicting the hypothesis that it wasn't. Hence $a \notin \langle p \rangle$, so $\langle p \rangle \subset \langle a \rangle$. If, however, $\langle a \rangle = R$, it would follow that a is a unit, again contradicting the hypothesis that it is noninvertible. So $\langle p \rangle \subset \langle a \rangle \subset R$, and hence $\langle p \rangle$ isn't a maximal principal ideal. Alternatively, p could be a unit, but then we would have $\langle p \rangle = R$, and again $\langle p \rangle$ wouldn't be a maximal principal ideal. Either way, $\langle p \rangle$ can't be a maximal principal ideal if p isn't irreducible.

(ii) Suppose $p \in R$ is prime and consider some $ab \in \langle p \rangle$. Then there exists some $c \in R$ for which $ab = cp$, so $p|ab$. Therefore either $p|a$, in which case $a = rp$ for some $r \in R$ and thus $a \in \langle p \rangle$, or $p|b$, whence $b = sp$ for some $s \in R$ and thus $b \in \langle p \rangle$. Hence $\langle p \rangle$ is a prime ideal.

Conversely, suppose $\langle p \rangle$ is a prime ideal, and $p = ab$ for some $a, b \in R$. Then $ab \in \langle p \rangle$, hence either $a \in \langle p \rangle$, in which case $p|a$, or $b \in \langle p \rangle$, whence $p|b$, both by Proposition 10.14. Thus p is prime in R . $\square$

Corollary 10.26 *Let R be an integral domain, and let p be some nonzero element of R . Then*

- (i) $R/\langle p \rangle$ is an integral domain if and only if p is prime, and
- (ii) if $R/\langle p \rangle$ is a field, then p is irreducible.

Proof (i) The ideal $\langle p \rangle$ is prime exactly when p is prime in R , and by Proposition 9.64 R/I is an integral domain exactly when I is a prime ideal of R .

(ii) By Proposition 9.61, if $R/\langle p \rangle$ is a field, then $\langle p \rangle$ is a maximal ideal. And since a maximal ideal that happens to be principal is necessarily a maximal principal ideal, p must be irreducible in R . $\square$

10.3 Principal ideal domains

Proposition 10.25 and Corollary 10.26 provide a neat connection between prime elements, prime ideals and integral domains as quotients, and a somewhat weaker and untidier link between irreducible elements, maximal ideals and fields as quotients.

As remarked just now, the problem arises when the integral domain under investigation has maximal ideals that aren't principal, and as we saw in Example 9.59 and Proposition 9.60 there are some by now quite familiar examples of integral domains that fall into this category.

Given two numbers which are relatively prime, find the least multiple of each such that one multiple exceeds the other by unity.

— Claude Gaspard Bachet de Méziriac
(1581–1638),
*Problemes plaisans et delectables, qui se
sont par les nombres* (1624) 18

There are, however, some integral domains that have only principal ideals, and when we're dealing with one of these, the second part of both Proposition 10.25 and Corollary 10.26 become much neater.

Definition 10.27 An integral domain R is said to be a **principal ideal domain** or **PID** if all of its ideals are principal; that is, of the form $\langle a \rangle$ for some $a \in R$.

Our model integral domain so far has been $\mathbb{Z}$, and it turns out that it all of its ideals are principal; that is, of the form $\langle n \rangle$ for some $n \in \mathbb{Z}$.

Proposition 10.28 *The ring $\mathbb{Z}$ is a PID.*

Proof The trivial ideal $\{0\} = \langle 0 \rangle$ is principal, as is $\mathbb{Z} = \langle 1 \rangle$ itself. Suppose that $I \subset \mathbb{Z}$ is a proper, nontrivial ideal, and choose the smallest nonzero element $b \in I$; that is, select $b \in I \setminus \{0\}$ such that $|b|$ is as small as possible. Then $\langle b \rangle \subseteq I$, since for any $n \in \mathbb{Z}$, the product nb lies in I by the absorption condition for ideals.

Now take some arbitrary element $a \in I$ and use the Division Theorem to find $q, r \in \mathbb{Z}$ such that $a = qb + r$, with either $r = 0$ or $|r| < |b|$. Then $r = 0$, because otherwise $0 < |r| < |b|$, which contradicts the choice of b . Hence $a = qb \in \langle b \rangle$, so $I \subseteq \langle b \rangle$ and therefore $I = \langle b \rangle$.

Thus any ideal of $\mathbb{Z}$ is principal, of the form $\langle b \rangle$ for some $b \in \mathbb{Z}$, and hence $\mathbb{Z}$ is a PID. $\square$

What this says is that every ideal of $\mathbb{Z}$ is generated by its smallest nonzero element, and is hence principal. So we can rewrite Proposition 10.25 and Corollary 10.26 in a nicer and stronger form in $\mathbb{Z}$. But actually, the above proof just makes use of basic properties of ideals and the fact that the Division Theorem works in $\mathbb{Z}$, so we should be able to extend it to any integral domain in which some version of the Division Theorem holds. This leads to the following, which is just a suitably generalised version of Proposition 10.28.

Proposition 10.29 *Every Euclidean domain is a PID.*

Proof Let R be a Euclidean domain, and let v be a Euclidean valuation for R . Then the trivial ideal $\{0\} = \langle 0 \rangle$ and $R = \langle 1 \rangle$ itself are principal. So, suppose that I is a proper, nontrivial ideal of R and choose $b \in I \setminus \{0\}$ such that $v(b)$ is as small as possible. Then $\langle b \rangle \subseteq I$.

Now let $a \in I$ be some arbitrary element of I . Because R is a Euclidean domain, we can find $q, r \in R$ such that $a = qb + r$, with either $r = 0$ or $v(r) < v(b)$. If $r \neq 0$ then $r = a - qb$, which is an element of I (since $a \in I$ and, by the absorption condition for ideals, $qb \in I$ as well). Then $v(r) < v(b)$, which can't be the case because we specifically chose b so that $v(b)$ was minimal over I . Therefore $r = 0$ and we have $a = qb \in \langle b \rangle$. So $I \subseteq \langle b \rangle$ and therefore $I = \langle b \rangle$ as claimed. Hence I is

principal, and R is a PID. $\square$

So, every Euclidean domain is a PID. But the converse doesn't necessarily hold: there are PIDs that aren't Euclidean. In particular, the ring $\mathbb{Z}[\theta]$, where $\theta = \frac{1}{2}(1 + \sqrt{-19})$, is a PID but not a Euclidean domain. The proof of this is a little involved and would be a bit of a digression at this point, so we postpone it to Section 10.A.

PIDs don't satisfy all the properties that Euclidean domains do: we're not guaranteed that some version of the Division Theorem will hold. Nevertheless, they are still quite civilised rings in which to work. In particular, we're guaranteed to have well-defined greatest common divisors and least common multiples:

Proposition 10.30 *If R is a PID then for any $a, b \in R$ there exists elements $d, l \in R$ such that $d = \gcd(a, b)$ and $l = \text{lcm}(a, b)$.*

Also, there exist $r, s \in R$ such that $ra + sb = d$.

The second part of this proposition is known as **Bézout's Identity**, after the French mathematician Etienne Bézout (1730–1783), who proved it for polynomials in his 1779 work *Théorie générale des équations algébriques*. The corresponding result for integers had been proved earlier by the French mathematician, poet and classicist Claude Gaspard Bachet de Méziriac (1581–1638) in his 1612 book *Problèmes plaisants et délectables qui se sont par les nombres*.

Proof Let $I = \langle a \rangle + \langle b \rangle = \{ra + sb : r, s \in R\}$. This is an ideal of R , and since R is a PID, I must be principal, and hence equal to $\langle d \rangle$ for some $d \in R$. This element d is a greatest common divisor of a and b : indeed, $\langle a \rangle \subseteq \langle d \rangle$ and $\langle b \rangle \subseteq \langle d \rangle$, so by Proposition 10.14 we have $d|a$ and $d|b$. Furthermore, for any $e \in R$ with $\langle a \rangle \subseteq \langle e \rangle$ and $\langle b \rangle \subseteq \langle e \rangle$, it follows that $\langle d \rangle = \langle a \rangle + \langle b \rangle \subseteq \langle e \rangle$, so $e|d$, and hence d is a greatest common divisor of a and b .

Since $\langle a \rangle + \langle b \rangle = \langle d \rangle$, there exists some element of $\langle a \rangle + \langle b \rangle$ equal to d . That is, there exist elements $r, s \in R$ such that $ra + sb = d$.

Similarly, $\langle a \rangle \cap \langle b \rangle$ is an ideal of R , and thus equal to $\langle l \rangle$ for some $l \in R$ since R is a PID. This l is a least common multiple of a and b : indeed, $\langle l \rangle \subseteq \langle a \rangle$ and $\langle l \rangle \subseteq \langle b \rangle$, so by Proposition 10.14 we have $a|l$ and $b|l$. Furthermore, for any $m \in R$ with $\langle m \rangle \subseteq \langle a \rangle$ and $\langle m \rangle \subseteq \langle b \rangle$, it follows that $\langle m \rangle \subseteq \langle a \rangle \cap \langle b \rangle = \langle l \rangle$ and thus $l|m$. Hence l is a least common multiple of a and b . $\square$

This can be extended to finite collections of elements:

Corollary 10.31 *Let R be a PID. Then for any $a_1, \dots, a_m \in R$ there exist elements $d, l \in R$ with $d = \gcd(a_1, \dots, a_m)$ and $l = \text{lcm}(a_1, \dots, a_m)$. Also, there exist elements $r_1, \dots, r_m \in R$ such that $r_1 a_1 + \dots + r_m a_m = d$.*



Wikimedia Commons / unknown artist

Claude Gaspard Bachet, Sieur de Méziriac (1581–1638) was born to a noble family in Bourg-en-Bresse, at the time part of the Duchy of Savoy. From early childhood, he lived and studied with the Jesuit Order, both his parents having died by the time he was six years old. He joined the order in 1601 but left the following year due to ill health and returned to his family estate at Bourg-en-Bresse, where he lived in comparative leisure for most of the rest of his life.

Bachet was an accomplished poet and writer, publishing poems in French, Italian and Latin, and translations of the psalms and other religious works, as well as translations of some of the works of Ovid (43BC–c.17AD). He was a member of the literary salon that in 1634 became the Académie Française, although recurring health problems prevented him from attending the inauguration ceremony and he was formally elected to membership the following year.

In addition to his literary works, he also wrote a number of books on mathematics, in particular *Problèmes plaisants et delectables qui se sont par les nombres* (1612). This book consisted of a collection of mathematical puzzles, mostly in the form of arithmetical problems, and was a forerunner of modern recreational mathematics books. He also published a Latin translation of the classic number theory text *Arithmetica* by the Greek mathematician Diophantus of Alexandria (c.200–c.285AD). It was in the margin of this edition that Pierre de Fermat (1601–1665) wrote his celebrated “Last Theorem”.



Wikimedia Commons / unknown artist

The son and grandson of lawyers, **Étienne Bézout** (1730–1783) instead pursued a career in mathematics, inspired by the work of the Swiss mathematician Leonhard Euler (1707–1783).

In the late 1750s he published memoirs on dynamics and integration, and in 1758 was elected an adjoint member of the Académie des Sciences. Over the next ten years he was appointed examiner to the Gardes de la Marine and the Corps d'Artillerie, with responsibility for the mathematical education of naval and army officer cadets.

In addition to his contributions to analysis, mechanics and algebra, he also wrote a number of influential textbooks, including the four-volume *Cours de mathématiques à l'usage des Gardes du Pavillon et de la Marine* (1764–1767) and the six-volume *Cours complet de mathématiques à l'usage de la marine et de l'artillerie* (1770–1782) which were used not only by officer cadets but also by students at the École Polytechnique, and in translation by Harvard and other American universities.

Bézout's approach to research, which he called the "method of simplifying assumptions" was to attack special cases of difficult problems, gradually developing greater insight that often enabled him to find a general solution.

What this says is not only that any two (or, actually, any finite collection of) elements of a PID have a greatest common divisor, but also that it can be written as a linear combination of those elements. We call integral domains with greatest common divisors **GCD domains**, and those that also satisfy the linearity condition are called **Bézout domains**, but we will not study them further here.

In Proposition 10.23 we saw that in an integral domain, every prime element is irreducible, but Example 10.24 showed that the converse doesn't necessarily hold. However, it does hold in a PID, as the next proposition shows.

Proposition 10.32 *In a PID R , every irreducible element is prime.*

Proof Suppose that $r \in R$ is irreducible, and that $r|ab$ for some $a, b \in R$. Then by Proposition 10.30 there exists an element $d \in R$ that is a greatest common divisor of r and a , and hence $r = ds$ for some $s \in R$. Since r is irreducible, either d or s is a unit. Suppose that d is a unit. By Proposition 10.30 we can find $x, y \in R$ such that $d = xa + yr$. Multiplying this by b we get $db = xab + yrb$. We know that $r|ab$ and obviously $r|yrb$, so $r|db$ as well, and hence $db = tr$ for some $t \in R$. But d is a unit, so we can multiply both sides by d^{-1} to get $b = d^{-1}tr$, which means that $r|b$.

If, on the other hand, s is a unit, we have $r \sim d$ and hence $r|d$. We know already that $d|a$, since it's a greatest common divisor of r and a , so therefore $r|a$.

Hence, if r is irreducible, either $r|a$ or $r|b$, and thus r is prime. $\square$

Combining this with Proposition 10.23 we immediately get the following fact.

Corollary 10.33 *In a PID, an element is irreducible if and only if it is prime.*

And putting it together with Proposition 10.25 we find that in PIDs there is a stronger link between prime and maximal ideals.

Corollary 10.34 *A nontrivial ideal of a PID is prime if and only if it is maximal.*

Proof Let R be a PID and let I be some nontrivial ideal of R . Then $I = \langle a \rangle$ for some $a \in R$ because R is a PID.

Then $I = \langle a \rangle$ is prime if and only if a is prime in R (by Proposition 10.25), a is prime in R exactly when it's irreducible in R (by Proposition 10.32), and a is irreducible in R if and only if $\langle a \rangle$ is a maximal principal ideal in R (by Proposition 10.25). But R is a PID, so maximal principal ideals and maximal ideals are the same thing. $\square$

So, in a PID irreducible elements and prime elements are the same, and prime ideals and maximal ideals are the same as well.

10.4 Unique factorisation domains

THE FUNDAMENTAL THEOREM OF ARITHMETIC says that any nonzero integer apart from ± 1 can be factorised uniquely (up to permutation and multiplication by ± 1) as a product of prime integers. Rephrasing this using the terminology of the last couple of sections, any nonzero, nonunit element of $\mathbb{Z}$ can be factorised as a product of irreducible elements. Also, any such factorisation in $\mathbb{Z}$ is unique up to permutation and multiplication by units. This leads to the following definition.

Definition 10.35 An integral domain R is a **factorisation domain** or **atomic domain** if any nonzero, nonunit element $r \in R$ can be expressed as a product $r = a_1 \dots a_m$ where $a_1, \dots, a_m$ are irreducible. Furthermore, R is a **unique factorisation domain** or **UFD** if for any two irreducible factorisations $r = a_1 \dots a_m = b_1 \dots b_n$, we have $m = n$, and there exists some permutation $\sigma \in S_n$ such that $a_i \sim b_{\sigma(i)}$ for all $1 \leq i \leq n$. That is, irreducible factorisations are unique up to permutation and multiplication by units.

Corollary 10.33 (which depends on Propositions 10.23 and 10.32) tells us that in a PID prime and irreducible elements are the same thing. This is true in UFDs as well:

Proposition 10.36 *Let R be a factorisation domain. Then R is a UFD if and only if every irreducible element is prime.*

Proof Let $x \in R$ be irreducible, suppose that R is a UFD, and suppose that $x|ab$ for some $a, b \in R$. We can factorise $a = r_1 \dots r_k$ and $b = r_{k+1} \dots r_m$, and hence $ab = r_1 \dots r_m$. But since $x|ab$ there exists some $y \in R$ such that $xy = ab$. Since R is a UFD we can factorise $y = s_1 \dots s_n$, and thereby obtain another factorisation $ab = xy = xs_1 \dots s_n$. But since R is a UFD, this factorisation must be unique up to permutation and association, and therefore $x \sim r_i$ for some i with $1 \leq i \leq m$. If $i \leq k$ then $x|a$, and if $i > k$ then $x|b$. Hence x is prime. This proves the first part: in a UFD, irreducible elements are prime.

Now for the converse. Suppose that R is at least a factorisation domain, and that every irreducible element in R is prime. Let $x \in R$ be neither zero nor a unit, and consider two irreducible factorisations $x = r_1 \dots r_m = s_1 \dots s_n$. We want to show that these are the same up to permutation and association (that is, multiplication by a unit). We

I never could do anything with figures, never had any talent for mathematics, never accomplished anything in my efforts at that rugged study, and to-day the only mathematics I know is multiplication, and the minute I get away up in that, as soon as I reach nine times seven ... I've got it now. It's eighty-four. Well, I can get that far all right with a little hesitation. After that I am uncertain, and I can't manage a statistic.

— Mark Twain (Samuel Langhorne Clemens) (1835–1910),
In Aid of the Blind (29 March 1906),
Mark Twain's Speeches (1910) 322–332

prove this by induction on m . If $m = 1$ then $x = r_1$ is irreducible, and hence $n = 1$ and $s_1 = r_1$.

Now suppose that any element $x \in R$ that can be expressed as a product of fewer than m irreducibles can be done so uniquely up to permutation and association. Since r_m divides $s_1 \dots s_n$, it must divide s_i for some $1 \leq i \leq n$. By permuting the factors, we can assume that $r_m | s_n$, which means that $s_n = qr_m$ for some $q \in R$. But q must be a unit, since both r_m and s_n are prime. So $r_m \sim s_n$, and hence $r_1 \dots r_m = s_1 \dots s_{n-1}(qr_m)$, which means that, by cancellation, $r_1 \dots r_{m-1} = qs_1 \dots s_{n-1}$. By the induction hypothesis we know that these two factorisations $r_1 \dots r_{m-1}$ and $qs_1 \dots s_{n-1}$ are the same up to permutation and multiplication by units. Hence R is a UFD. $\square$

Let's now summarise what we know about the relationship between primes and irreducibles in integral domains.

- In an integral domain, all primes are irreducible.⁸
- In a UFD, all irreducible elements are prime.⁹
- In a PID, all irreducible elements are prime.¹⁰
- A factorisation domain where all irreducible elements are prime is a UFD.¹¹

We saw earlier that every Euclidean domain is a PID.¹² By the above discussion, every PID that is also a factorisation domain is also a UFD. So if we could show that every PID is a factorisation domain, we'd have a nice, neat chain of implications: a Euclidean domain is a PID, which is a UFD. To settle this question, we'll look at factorisation in $\mathbb{Z}$.

Proposition 10.37 *Any nonzero integer n apart from ± 1 can be expressed as a finite product $n = p_1 \dots p_k$ of primes. That is, $\mathbb{Z}$ is a factorisation domain.*

Proof Suppose n can't be factorised as a product of primes. Then n itself can't be prime, and we can express it as a product $n = n_{1,1}n_{1,2}$ where neither $n_{1,1}$ nor $n_{1,2}$ is ± 1 . At least one (and possibly both) of these is not prime, otherwise $n = n_{1,1}n_{1,2}$ would be a prime factorisation. Suppose, without loss of generality, that this non-prime integer is $n_{1,1}$, and let $n_1 = n_{1,1}$. Then we can factorise it further as $n_1 = n_{1,1} = n_{2,1}n_{2,2}$. At least one of these factors, say $n_{2,1}$, must be prime, so let $n_2 = n_{2,1}$ and factorise it as $n_{3,1}n_{3,2}$. We continue this process to obtain an infinite sequence $n = n_0, n_1, n_2, \dots$ of integers such that $n_{i+1} | n_i$ but $n_i \neq \pm n_{i+1}$.

But n is finite, so this sequence can't be infinite and must eventually terminate. Thus n must be factorisable as a finite product of primes. $\square$

The following corollary is known as the **Fundamental Theorem of Arithmetic**.¹³

⁸ Proposition 10.23, page 389.

⁹ Proposition 10.36, page 395.

¹⁰ Proposition 10.32, page 394.

¹¹ Proposition 10.36, page 395.

¹² Proposition 10.29, page 392.

¹³ The earliest known proof appears as Proposition 32 of Book VII of Euclid's *Elements*; a more modern proof appears in Gauss' 1801 book *Disquisitiones Arithmeticae*.

Corollary 10.38 (Fundamental Theorem of Arithmetic) *The ring $\mathbb{Z}$ of integers is a UFD. That is, any nonzero integer apart from ± 1 can be factorised as a finite product of prime (and hence irreducible) integers. Furthermore, any such factorisation is unique up to permutation and multiplication of individual factors by ± 1 .*

Proof Proposition 10.37 says that $\mathbb{Z}$ is a factorisation domain, and by Proposition 10.28 we know $\mathbb{Z}$ is a PID. Hence $\mathbb{Z}$ must be a UFD. $\square$

We'd like to generalise the above argument to other integral domains. Notice that the key ingredient in the proof was showing that for any nonzero, nonunit element in $\mathbb{Z}$ we can find a finite sequence $n = n_0, n_1, \dots, n_k$ of nonunit elements such that $n_{i+1} | n_i$ but $n_{i+1} \not\sim n_i$. By Proposition 10.14 this is equivalent to requiring the existence of a finite sequence of principal ideals

$$\langle n_k \rangle \subset \langle n_{k-1} \rangle \subset \dots \subset \langle n_1 \rangle \subset \langle n_0 \rangle = \langle n \rangle \subset \mathbb{Z}$$

for any nonzero nonunit integer $n \in \mathbb{Z}$.

So, any integral domain R which only has finite chains of strictly-included principal ideals should be a factorisation domain by some suitably modified version of the above proof. Let's introduce some new terminology and then prove this more general result.

Definition 10.39 If a ring R contains no infinite, strictly increasing chain of principal ideals

$$\langle a_1 \rangle \subset \langle a_2 \rangle \subset \dots \subset R$$

then we say it satisfies the **ascending chain condition for principal ideals**.

In the rest of this section we will be interested mainly in integral domains that satisfy the ACCP, but it's worth also mentioning the following stronger condition:

Definition 10.40 A ring R is said to be **Noetherian** if it satisfies the **ascending chain condition (ACC)**: if every strictly increasing chain of (not necessarily principal) ideals

$$I_1 \subset I_2 \subset \dots \subset R$$

terminates after finitely many steps.

More generally, R is **left Noetherian** if it satisfies the ACC for left ideals, and **right Noetherian** if it satisfies the ACC for right ideals.

Many of the rings we've met so far are Noetherian. Examples of rings that fail to satisfy the ACC include $\mathbb{F}[x_1, x_2, \dots]$, the polynomial ring with infinitely many unknowns, which contains infinite chains of the form $\langle x_1 \rangle \subset \langle x_1, x_2 \rangle \subset \langle x_1, x_2, x_3 \rangle \subset \dots$, and the ring of **algebraic integers** (that is, those complex numbers that are roots of polynomials

in $\mathbb{Z}[x]$ with leading coefficient 1), which contains infinite chains of the form $\langle 2 \rangle \subset \langle 2^{1/2} \rangle \subset \langle 2^{1/3} \rangle \subset \langle 2^{1/4} \rangle \subset \dots$.

What we want to do now is to generalise the proof of Proposition 10.37 to the case of any integral domain satisfying the ACC for principal ideals.

Proposition 10.41 *Let R be an integral domain that satisfies the ACC for principal ideals. Then R is a factorisation domain.*

Proof For the moment, we'll call a nonzero, nonunit element $r \in R$ **unfactorisable** if it can't be factorised as a finite product of irreducible elements of R . Since r isn't irreducible, it can be expressed as a product $r = a_1 b_1$, where neither a_1 nor b_1 are units. At least one of these must be unfactorisable; suppose without loss of generality that a_1 is this element. Then a_1 can't be irreducible, so it can be expressed as a product $a_1 = a_2 b_2$, at least one of which, say a_2 , must be unfactorisable.

If r were unfactorisable, then this process must continue forever, yielding an infinite sequence $r = a_0, a_1, a_2, \dots$ of unfactorisable elements of R . Furthermore, $a_{i+1} | a_i$ but $a_i \not\sim a_{i+1}$ for all $i \geq 0$. So by Proposition 10.14 we get an infinite chain

$$\langle r \rangle = \langle a_0 \rangle \subset \langle a_1 \rangle \subset \langle a_2 \rangle \subset \dots$$

of principal ideals in R .

But this can't happen, because R satisfies the ACC for principal ideals. Hence every element of R must be factorisable as a product of irreducible elements, and therefore R is a factorisation domain. $\square$

All that remains is to show that PIDs satisfy the ACC for principal ideals.

Proposition 10.42 *Let R be a PID. Then R satisfies the ACC for principal ideals.*

Proof Suppose that $I_1 \subset I_2 \subset \dots$ is a strictly increasing chain of ideals in R . We want to show that this must be of finite length.

The union $I = \bigcup_i I_i = I_1 \cup I_2 \cup \dots$ is itself an ideal of R . To confirm this, we need to check that it satisfies the conditions for a (not necessarily unital) subring, and also obeys the absorption condition.

It certainly contains 0, since each of the I_i do. For any $a \in I_k \subseteq I$ there exists $-a_k \in I_k \subseteq I$. Now suppose that $a \in I_j$ and $b \in I_k$ for some $j, k \geq 1$. Either $j = k$, in which case $a, b \in I_j$ and hence both $a+b$ and ab also lie in $I_j \subseteq I$, or $j \neq k$, in which case we may assume without loss of generality that $j < k$, and hence $a \in I_j \subset I_k$, so $a+b$ and ab lie in I_k . Thus I is a (not necessarily unital) subring of R .

Now suppose that $r \in R$ and $a \in I$. Then $a \in I_k \subseteq I$ for some $k \geq 1$,

and since I_k is an ideal of R it follows that both ar and ra lie in $I_k \subseteq I$. So I is an ideal of R .

Since R is a PID, $I = \langle a \rangle$ for some $a \in R$. This element a must belong to some ideal I_n in the chain, and so for any other ideal I_k we must have $I_k \subseteq I = \langle a \rangle = I_n$. Therefore I_n must be the last ideal in the chain, and so R satisfies the ACC for principal ideals. $\square$

Putting all this together, we have the following corollary:

Corollary 10.43 *Every PID is a UFD.*

Proof We've just shown that every PID satisfies the ACC for principal ideals, and is hence a factorisation domain. By Proposition 10.32, in a PID all irreducible elements are prime, and by Proposition 10.36 every factorisation domain where all irreducible elements are prime is a UFD. Therefore every PID is a UFD. $\square$

In Proposition 10.30 we saw that greatest common divisors and least common multiples exist in PIDs. This is also true in UFDs:

Proposition 10.44 *Let R be a UFD, and let a and b be two nonzero elements of R . Then there exist $d, l \in R$ such that d is a greatest common divisor of a and b , and l is a least common multiple of a and b .*

Proof Since R is a UFD, both a and b have finitely many irreducible factors. Let $p_1, \dots, p_k$ be the irreducible elements that divide either a or b , such that none is an associate of any of the others. That is, we make a complete list of all the irreducible factors of a , and all the irreducible factors of b , and choose only one representative from each equivalence class of associate elements, discarding the others. We can then write

$$a = up_1^{m_1} \dots p_k^{m_k} \quad \text{and} \quad b = vp_1^{n_1} \dots p_k^{n_k},$$

where the exponents $m_1, \dots, m_k, n_1, \dots, n_k \geq 0$ are nonnegative integers, and $u, v \in R$ are units.

Then $d = p_1^{e_1} \dots p_k^{e_k}$ and $l = p_1^{f_1} \dots p_k^{f_k}$ are the required elements, where $e_i = \min(m_i, n_i)$ and $f_i = \max(m_i, n_i)$ for $1 \leq i \leq k$. $\square$

This proof can be extended in a straightforward manner to finite collections of nonzero elements:

Corollary 10.45 *Let R be a UFD, and let $a_1, \dots, a_n \in R$ be nonzero elements. Then there exist $d, l \in R$ such that d is a greatest common divisor and l is a least common multiple of $a_1, \dots, a_n$.*

So, greatest common divisors and least common multiples exist in a UFD. However, we haven't shown that greatest common divisors satisfy Bézout's identity; that is, that they can be written as linear combinations of the given elements. Indeed, this is not the case: there

exist UFDs that don't satisfy Bézout's condition, although we won't study these any further.

For the moment, however, we'll concern ourselves with finding examples of UFDs that aren't PIDs. Some important examples are given by univariate polynomial rings $R[x]$ for some carefully chosen ring R . To that end, we'll now spend a little time studying factorisation in polynomial rings. This will also serve as a prelude to the material in the next chapter.

First of all, we consider the case where R is an integral domain.

Proposition 10.46 *If R is an integral domain, then so is $R[x]$.*

Proof We know already that $R[x]$ is a ring, and if R is commutative (which as an integral domain, it must be) then $R[x]$ must be too. Also, the unity element $1 \in R$ also serves as the unity element in $R[x]$. All that remains is to show that $R[x]$ contains no zero divisors. To this end, let $f = a_mx^m + \cdots + a_1x + a_0$ and $g = b_nx^n + \cdots + b_1x + b_0$ be two elements of $R[x]$, with both a_m and b_n nonzero (so $\deg(f) = m$ and $\deg(g) = n$). Then fg has leading term $a_mb_nx^{m+n}$, which can only be zero if $a_mb_n = 0$. But we know this isn't the case, since R is an integral domain, and hence $R[x]$ can have no zero divisors either, and must therefore be an integral domain. $\square$

Univariate polynomial rings over fields satisfy stronger conditions:

Proposition 10.47 *If $\mathbb{F}$ is a field, then $\mathbb{F}[x]$ is a PID. Furthermore, a nontrivial ideal I in $\mathbb{F}[x]$ is equal to $\langle g \rangle$ if and only if g is a nonzero polynomial of minimal degree in I .*

Proof We know from Proposition 10.46 that $\mathbb{F}[x]$ must at least be an integral domain, so to prove that it's a PID we just need to show that every ideal of $\mathbb{F}[x]$ is principal.

Suppose, then, that I is an ideal of $\mathbb{F}[x]$. If I is trivial, then $I = \{0\} = \langle 0 \rangle$. Otherwise, choose a nonzero polynomial $g \in I$ with degree less than any other nonzero element of I . Certainly $\langle g \rangle \subseteq I$. By Theorem 10.5 we can decompose any other element $f \in I$ as $f = qg + r$, where $r = 0$ or $\deg(r) < \deg(g)$.

Then $r = f - qg$, and since $f, g \in I$ it follows that $r \in I$. If $r \neq 0$ then $\deg(r) < \deg(g)$, which contradicts our hypothesis that g has minimal degree in I . So we must have $r = 0$, which means that $f = qg \in \langle g \rangle$. So $I \subseteq \langle g \rangle$ and hence $I = \langle g \rangle$. $\square$

We've now considered two, in some sense, extreme cases of $R[x]$: one when R is just an integral domain, and the other when R is a field. We now want to look at the intermediate case where R is a UFD. To do this, we have to think carefully about irreducible elements in $R[x]$.

Proposition 10.48 *Let R be a integral domain. The units in $R[x]$ are exactly the units in R , and an irreducible element in R is also irreducible in $R[x]$.*

Proof Suppose that $f, g \in R[x]$ with $f = a_mx^m + \cdots + a_1x + a_0$ and $g = b_nx^n + \cdots + b_1x + b_0$ and that $fg = 1$. Then

$$fg = a_mb_nx^{m+n} + \cdots + a_0b_0 = 1,$$

and by comparing coefficients we find that

$$a_1 = \cdots = a_m = b_1 = \cdots = b_n = 0$$

and $a_0b_0 = 1$, which means that a_0 and b_0 must be units in both R and $R[x]$. Conversely, if $a, b \in R$ are units, with $ab = 1$, then this also holds in $R[x]$. Hence the units in $R[x]$ are exactly the units in R .

Now suppose that $r \in R$ is irreducible in R . Suppose that $r = ab$ where neither a nor b are units in R . By the argument above, neither a nor b can be units in R either. Since r is irreducible in R , at least one of a and b must lie in $R[x]$ but not in R ; that is, they must be non-constant polynomials of strictly positive degree. Suppose, without loss of generality, that $\deg(a) = m > 0$. Then

$$\deg(r) = \deg(ab) = \deg(a) + \deg(b) = m + \deg(b) > 0.$$

But $r \in R$ is a constant polynomial in $R[x]$, and so $\deg(r) = 0$, which contradicts this. Therefore r must be irreducible in $R[x]$. $\square$

The second part of this proposition says irreducibles in R must also be irreducible in $R[x]$, but it doesn't rule out the possibility that there might be other irreducible elements in $R[x]$ apart from the ones inherited from R . To properly understand the irreducible elements in $R[x]$ we need to study a particular class of polynomials: those that can't be expressed as a scalar multiple of another polynomial.

Definition 10.49 Let R be a UFD. A nonzero polynomial

$$f = a_mx^m + \cdots + a_1x + a_0 \in R[x]$$

is **primitive** if $\gcd(a_0, \dots, a_m) = 1$; that is, if its coefficients are coprime. The element $c(f) = \gcd(a_0, \dots, a_m)$ is the **content** of f .

Somewhat confusingly, the term **primitive polynomial** has another, different meaning in the context of finite fields. We'll discuss this further in the next chapter.

In particular, this means that any nonzero polynomial

$$f = a_mx^m + \cdots + a_1x + a_0 \in R[x]$$

can be written as $f = af_0$, where f_0 is a primitive polynomial in $R[x]$, and $a = c(f) = \gcd(a_0, \dots, a_m)$.

Example 10.50 The polynomials

$$3x^2 + x - 2, \quad 2x + 1 \quad \text{and} \quad x^5 - 2x + 4$$

are primitive in $\mathbb{Z}[x]$, but

$$2x^2 + 4x + 2, \quad 3x - 9 \quad \text{and} \quad 4x^5 - 8x^4 - 2x^3$$

aren't.

We'll see in a little while that primitive polynomials yield all the other irreducible elements in $R[x]$. For the moment, however, we'll look at what happens when we multiply two primitive polynomials. Let $f = 3x^2 + x - 2$, and let $g = x^5 - 2x + 4$. These are both primitive in $\mathbb{Z}[x]$. Then

$$fg = 3x^7 + x^6 - 2x^5 - 6x^3 + 10x^2 + 8x - 8.$$

This is also primitive in $\mathbb{Z}[x]$, since its coefficients are all coprime. The fact that this holds in general for polynomials in $\mathbb{Z}[x]$ was originally stated and proved by Carl Friedrich Gauss (1777–1855).¹⁴ The following is a generalisation of Gauss' original statement to any UFD.

¹⁴ It appears as Section 42 in Chapter 2 (*De Congruentiis Primi Gradus*) of his book *Disquisitiones Arithmeticae* (1801).

Proposition 10.51 (Gauss' Lemma) *Let R be a UFD, and let*

$$f = a_mx^m + \cdots + a_1x + a_0 \quad \text{and} \quad g = b_nx^n + \cdots + b_1x + b_0$$

be two primitive polynomials in $R[x]$. Then the product fg is also primitive in $R[x]$.

Proof Suppose that the product $fg = c_{m+n}x^{m+n} + \cdots + c_1x + c_0$ is not primitive; that is, there exists some irreducible element $p \in R$ that divides all the coefficients c_i for $0 \leq i \leq m+n$.

Both f and g are primitive, so this element p can't divide all of their coefficients. In that case, there must be an integer $s \geq 0$ such that $p|a_i$ for $i < s$ but $p \nmid a_s$; similarly there must be an integer $t \geq 0$ such that $p|b_i$ for $i < t$ but $p \nmid b_t$. That is, s and t are, respectively, the indices of the lowest-degree coefficients of f and g that are not divisible by p .

The coefficients of the product fg are completely determined by the coefficients of f and g , with $c_k = \sum_{i+j=k} a_ib_j$ for $0 \leq k \leq m+n$. Since p is irreducible and R is a UFD, by Proposition 10.36 p must therefore be prime, and thus if $p \nmid a_s$ and $p \nmid b_t$, it must also be the case that $p \nmid a_sb_t$. Now consider the coefficient $c_{s+t} = \sum_{i+j=s+t} a_ib_j$. The term a_sb_t in this sum is not divisible by p , but all of the other terms a_ib_j are, because either $i < s$ (and hence $p|a_i$) or $j < t$ (so $p|b_j$).

Thus p can't divide c_{s+t} , which contradicts our original statement that p does divide all the coefficients c_i of fg . Hence no such irreducible element p can exist, and fg must also be primitive. $\square$

There is a short and elegant proof of the special case where $R = \mathbb{Z}$, which makes use of the fact that $\mathbb{Z}_p[x]$ is an integral domain for any prime integer p .

Proposition 10.52 *Let $f, g \in \mathbb{Z}[x]$ be primitive polynomials. Then fg is also primitive in $\mathbb{Z}[x]$.*

Proof Suppose that the product fg is not primitive, so there exists some prime integer p that divides all of its coefficients. There is a well-defined homomorphism $\phi_p: \mathbb{Z} \rightarrow \mathbb{Z}_p$ mapping every integer n to its residue $[n]_p$ modulo p , and we can extend this to a ring homomorphism $\hat{\phi}_p: \mathbb{Z}[x] \rightarrow \mathbb{Z}_p[x]$ that maps a given polynomial in $\mathbb{Z}[x]$ to the corresponding polynomial in $\mathbb{Z}_p[x]$ with coefficients reduced modulo p .

Then $\hat{\phi}_p(f) \neq 0$ and $\hat{\phi}_p(g) \neq 0$, since p doesn't divide all the coefficients of either f or g . However, p does divide all the coefficients of their product fg , so $\hat{\phi}_p(fg) = 0$.

Since $\hat{\phi}_p$ is a ring homomorphism, we have

$$\hat{\phi}_p(f)\hat{\phi}_p(g) = \hat{\phi}_p(fg) = 0.$$

But this can't happen, because $\mathbb{Z}_p[x]$ is an integral domain by Proposition 10.46, and therefore has no zero divisors. Hence fg must be primitive in $\mathbb{Z}[x]$. $\square$

We know from Proposition 10.48 that the irreducible elements of R must also be irreducible in $R[x]$: these are exactly the irreducible constant (degree-0) polynomials in $R[x]$. We now want to find all the irreducible non-constant polynomials. These must be primitive, because otherwise they could be factorised as a product of a primitive polynomial and a non-unit constant element. But not all primitive polynomials are irreducible, because Gauss' Lemma (Proposition 10.51) says that a product of primitive polynomials is itself primitive. Thus there must exist some primitive polynomials that can be factorised, and we'd like to find out, in some sense, exactly which ones these are.

The key, it turns out, involves the field of fractions $Q(R)$ of our chosen UFD R .¹⁵ The irreducible polynomials in $R[x]$ are exactly the ones which are also irreducible in $Q[x]$. To see this, we'll look at an illustrative example first, and then use those ideas to prove it in general.

¹⁵ Definition 8.51, page 337.

It's relatively straightforward to show that a reducible polynomial in $\mathbb{Z}[x]$ must also be reducible in $\mathbb{Q}[x]$.

Example 10.53 Let $f = 2x^3 - 6x^2 + x - 3$ in $\mathbb{Z}[x]$. This is primitive, since $c(f) = 1$, but factorises as $f = (x - 3)(2x^2 + 1)$ and is thus reducible in $\mathbb{Z}[x]$. It must also be reducible in $\mathbb{Q}[x]$, since its factors $g = x - 3$ and $h = 2x^2 + 1$ also lie in $\mathbb{Q}[x]$.

The only thing that would stop a primitive reducible polynomial $f = gh \in \mathbb{Z}[x]$ from being reducible in $\mathbb{Q}[x]$ as well, would be if one

of its factors g or h happened to be a unit in $\mathbb{Q}[x]$ but not in $\mathbb{Z}[x]$. But Proposition 10.48 says that the units in $\mathbb{Q}[x]$ are exactly the units in $\mathbb{Q}$, namely the nonzero rationals, and the primitivity of f means that neither g nor h can be a constant in $\mathbb{Z} \subset \mathbb{Q}$: they must both have degree at least 1.

Going the other way is slightly more involved. We'll use the same polynomial as in the previous example to illustrate the technique we're going to use in the proof.

Example 10.54 Let $f = 2x^3 - 6x^2 + x - 3$ in $\mathbb{Q}[x]$, and also in $\mathbb{Z}[x]$. This can be factorised as $f = (x^2 + \frac{1}{2})(2x - 6)$ and is hence reducible in $\mathbb{Q}[x]$. We want to find a factorisation into non-units in $\mathbb{Z}[x]$, but this one won't work, because $g = (x^2 + \frac{1}{2})$ isn't in $\mathbb{Z}[x]$, even though the other factor $h = (2x - 6)$ is.

We start by getting rid of any awkward fractions. We do this by looking at the least common multiple of the denominators of the coefficients of a given factor, and then multiplying through by that number. In the case of g we get $a_1 = \text{lcm}(1, 2) = 2$, and for h we get $b_1 = \text{lcm}(1, 1) = 1$. Multiplying through, we get $a_1g = 2x^2 + 1$ and $b_1h = 2x - 6$.

These can now be considered as elements of $\mathbb{Z}[x]$, but the problem now is that b_1h isn't primitive. So we fix that by dividing by the content. Let $a_2 = c(a_1g) = 1$ and $b_2 = c(b_1h) = 2$, and then set $a = \frac{a_1}{a_2} = 2$ and $b = \frac{b_1}{b_2} = \frac{1}{2}$.

Now both $ag = 2x^2 + 1$ and $bh = x - 3$ are primitive elements of $\mathbb{Z}[x]$, and it so happens that $(ag)(bh) = (2x^2 + 1)(x - 3) = f$ is the required factorisation of f in $\mathbb{Z}[x]$, hence f is reducible in $\mathbb{Z}[x]$ as well as in $\mathbb{Q}[x]$.

In the proof of the next proposition, we'll see that we can always do this: for any primitive polynomial in $\mathbb{Z}[x]$ that factorises into lower-degree polynomials in $\mathbb{Q}[x]$, we can find a corresponding factorisation in $\mathbb{Z}[x]$ too.

Proposition 10.55 *Let R be a UFD, and let $Q = Q(R)$ be its field of fractions. Then a primitive polynomial $f \in R[x]$ is irreducible in $R[x]$ if and only if it is irreducible in $Q[x]$.*

Proof If $\deg(f) = 0$ then f is a constant element of $R[x]$ and also a unit in R . It is therefore not irreducible in $R[x]$ or in $Q[x]$.

So, we must have $\deg(f) \geq 1$. Suppose that $f = gh$ for some nonunit polynomials in $R[x]$. Since f is primitive, neither g nor h can be constants, so neither $\deg(g)$ nor $\deg(h)$ can be zero. By Proposition 10.48 neither g nor h can be units in $\mathbb{Q}[x]$ either, and so f must be reducible

in $Q[x]$. Contrapositively, if f is irreducible in $Q[x]$, it must be irreducible in $R[x]$ too.

Now suppose that f is primitive in $R[x]$ and reducible in $Q[x]$. Then $f = gh$ with $g, h \in Q[x]$ not units, so $\deg(g)$ and $\deg(h)$ are positive. Let a_1 be a least common multiple of the denominators of the coefficients of g , so that a_1g has coefficients in R , and thus $a_1g \in R[x]$. Next, let $a_2 = c(a_1g)$, the content of a_1g . Then $a = \frac{a_1}{a_2}$ is an element of Q , and ag is a primitive polynomial in $R[x]$.

Similarly, let b_1 be a least common multiple of the denominators of the coefficients of h , and let $b_2 = c(b_1h)$. Then if $\frac{b_1}{b_2}$, we have $b \in Q$ and bh a primitive element of $R[x]$.

By Gauss' Lemma,¹⁶ $k = (ag)(bh) = abgh = abf$ is primitive in $R[x]$,¹⁶ Proposition 10.51, page 402. so $a_1b_1f = a_2b_2k$.

Hence a_1b_1 and a_2b_2 are both greatest common divisors of the coefficients of a_1b_1f , and therefore $a_1b_1 \sim a_2b_2$ in R . But this means that $u = ab = \frac{a_1b_1}{a_2b_2}$ is a unit in R , and hence $f = (ag)(u^{-1}bh)$ is a factorisation of f in $R[x]$. So f must also be reducible in $R[x]$.

Contrapositively, if f is irreducible in $R[x]$ it must also be irreducible in $Q[x]$. $\square$

The special case where $R = \mathbb{Z}$ and $Q = Q(\mathbb{Z}) = \mathbb{Q}$ was originally proved by Gauss, and is also sometimes called Gauss' Lemma.

We are now able to completely characterise irreducible elements of $R[x]$ where R is a UFD:

Corollary 10.56 *Let R be a UFD, and let $Q = Q(R)$ be its field of fractions. Then an irreducible element in $R[x]$ is either irreducible in R , or a primitive polynomial in $R[x]$ that is also irreducible in $Q[x]$.*

Proof We know from Proposition 10.48 that an irreducible element of R is irreducible in $R[x]$. By Proposition 10.55 any primitive polynomial in $R[x]$ is irreducible in $R[x]$ if and only if it is irreducible in $Q[x]$. Finally, any non-primitive polynomial in $R[x]$ is reducible in $R[x]$. $\square$

So, the irreducible elements in $R[x]$ comprise the irreducible constants inherited from R , and the primitive polynomials in $R[x]$ that are also irreducible in $Q[x]$.

We now have all the ingredients we need to prove the last major result of this section:

Proposition 10.57 *If R is a UFD, then so is $R[x]$.*

Proof We must show that every nonzero nonunit in $R[x]$ has a factorisation into irreducible elements, and then that this factorisation is unique up to permutation and association.

The first of these is intuitively obvious, but we will prove it in detail anyway. Let $f \in R[x]$ be nonzero and not a unit. Then if $\deg(f) = 0$ it follows that $f \in R$ and hence has a unique factorisation because R is a UFD.

So, from now on, suppose that $\deg(f) > 0$. Regard f as a polynomial in $Q[x]$, where $Q = Q(R)$ is the field of fractions of R . Then since Q is a field, $Q[x]$ is a Euclidean domain by Theorem 10.5 and hence a PID by Proposition 10.29 and thus a UFD by Corollary 10.43. So f has a unique factorisation

$$f = g_1 \cdots g_k$$

in $Q[x]$, where $g_1, \dots, g_k$ are irreducible in $Q[x]$.

The coefficients of all these irreducible polynomials are of the form $\frac{a}{b}$ with $a, b \in R$, so next we clear the denominators in the same way as in the proof of Proposition 10.55. Let a_i be a least common multiple of the denominators of the coefficients of g_i , for $1 \leq i \leq k$, and let $a = a_1 \cdots a_k$. Then

$$af = h_1 \cdots h_k$$

where $a \in R$ and $h_1, \dots, h_k \in R[x]$.

Now $g_1, \dots, g_k$ were all irreducible in $Q[x]$, so $h_1, \dots, h_k$ are all irreducible in $Q[x]$ as well, because each h_i is just the corresponding g_i multiplied by some nonzero element of R that must therefore be a unit in Q .

Considering the content of each of the polynomials $f, h_1, \dots, h_k$, we get

$$f = cr \quad \text{and} \quad h_i = c_i s_i,$$

for $1 \leq i \leq k$, where $r, s_1, \dots, s_k \in R[x]$ are primitive, $c = c(f)$ and $c_i = c(h_i)$. Then

$$\begin{aligned} (ac)r &= (c_1 s_1) \cdots (c_k s_k) \\ &= (c_1 \cdots c_k) s_1 \cdots s_k. \end{aligned}$$

The content of a polynomial is a least common multiple, and hence is unique up to association, so there must be some unit $u \in R$ such that

$$acu = c_1 \cdots c_k.$$

Then

$$(ac)r = (acu) s_1 \cdots s_k$$

and hence

$$f = cr = (cu) s_1 \cdots s_k.$$

Now cu has a unique factorisation as irreducible elements of $R \subset R[x]$ since R is a UFD. Furthermore $s_1, \dots, s_k$ are all irreducible in $R[x]$ because they were primitive in $R[x]$ and were carefully selected to

be irreducible in $Q[x]$, so by Proposition 10.55 they are irreducible in $R[x]$.

Therefore f has a factorisation into irreducible elements in $R[x]$ and hence $R[x]$ is a factorisation domain.

Now we want to show that such a factorisation is unique up to permutation and association. Suppose that

$$f = p_1 \dots p_k g_1 \dots g_m = q_1 \dots q_l h_1 \dots h_n$$

are factorisations of the polynomial $f \in R[x]$, with $p_1, \dots, p_k, q_1, \dots, q_l$ irreducible elements of R , and $g_1, \dots, g_m, h_1, \dots, h_n$ irreducible primitive polynomials in $R[x]$.

By Proposition 10.55, $g_1, \dots, g_m, h_1, \dots, h_n$ are irreducible in $Q[x]$, and since $Q[x]$ is a UFD, $m = n$ and, possibly after some judicious permutation, $g_i \sim h_i$ for $1 \leq i \leq m$. This means that $h_i = u_i g_i$ for some units $u_1, \dots, u_m \in Q$. Since Q is the field of fractions of R , each $u_i = \frac{a_i}{b_i}$ for some $a_i, b_i \in R$, and thus $b_i h_i = a_i g_i$. Since h_i and g_i are both primitive, a_i and b_i are both greatest common divisors of the coefficients of $a_i g_i$, so $b_i \sim c_i$ in R , and hence each u_i is a unit in R as well as in Q .

We can now cancel all the g_i to get

$$u p_1 \dots p_k = q_1 \dots q_l$$

where $u = u_1 \dots u_m$ is a unit in R . Since R is a UFD, it follows that $k = l$ and, after a suitable permutation, $p_i \sim q_i$ for $1 \leq i \leq k$. Hence the factorisation of f is unique up to permutation and association, and therefore $R[x]$ is a UFD as claimed. $\square$

We can extend this result to multivariate polynomial rings by induction:

Corollary 10.58 *If R is a UFD then so is $R[x_1, \dots, x_n]$.*

Proof This follows by induction on n , using the observation that $R[x][y] \cong R[x, y]$. That is, the ring of polynomials in the formal variable y with coefficients from $R[x]$, is isomorphic to the ring of polynomials in the formal variables x and y with coefficients in R . $\square$

We are now, finally, in a position to show that although every PID is a UFD, not every UFD is a PID.

Example 10.59 The polynomial ring $\mathbb{Z}[x]$ is a UFD by Proposition 10.57, since $\mathbb{Z}$ is a UFD. But $\mathbb{Z}[x]$ is not a PID because, for example, the ideal $\langle x, 2 \rangle$ is not principal.

This almost completes the verification of the strictness in this chapter's hierarchy

$$\{\text{Euclidean domains}\} \subset \{\text{PIDs}\} \subset \{\text{UFDs}\} \subset \{\text{integral domains}\}$$

of integral domains. Not every integral domain is a UFD: the ring $\mathbb{Z}[\sqrt{-3}]$ in Example 10.24 is an integral domain but doesn't have unique factorisation because $4 = 2 \times 2 = (1 + \sqrt{-3})(1 - \sqrt{-3})$. We've just shown that not every UFD is a PID. It also so happens that not every Euclidean domain is a PID, but probably the simplest counterexample is the ring $\mathbb{Z}[\alpha]$ where $\alpha = \frac{1}{2}(1 + \sqrt{-19})$, an example of a **quadratic integer ring**. The proof that it is a PID but not a Euclidean domain is a bit involved, and we leave it to Section 10.A.

We will return to discussion of irreducible polynomials in the next chapter, which is a short overview of Galois theory, an important branch of mathematics which applies powerful techniques in group and ring theory to the solubility of polynomial equations.

Therefore, the Divine Spirit discovered an elegant and wonderful escape in that miracle of analysis, that token of the ideal world, that amphibian between being and non-being, that we call the imaginary root.

— Gottfried Wilhelm Leibniz
(1646–1716),

*Specimen novum analyseos pro scientia
infiniti, circa summas et quadraturas*
(1702)

10.A Quadratic integer rings

IN THIS SECTION, WE WILL STUDY THE GAUSSIAN INTEGERS $\mathbb{Z}[i]$ and investigate which of them are prime. We will then extend some of these ideas to a class of rings called **quadratic integer rings**, and finally construct an example of a PID that isn't a Euclidean domain.

We saw in Proposition 10.11 that $\mathbb{Z}[i]$ is a Euclidean domain, so irreducible and prime elements are the same thing. The Euclidean valuation we used before was defined by

$$v(a + bi) = |a + bi|^2 = a^2 + b^2.$$

(It's also worth noting that $v(a + bi) = (a + bi)\overline{(a + bi)}$.) In a little while we will develop a general technique for defining suitable valuation functions, but for the moment this will suffice. In particular, we want to see how this valuation behaves with respect to primes in $\mathbb{Z}$ or $\mathbb{Z}[i]$. The first thing to notice is that this valuation satisfies a multiplicative property:

$$\begin{aligned} v((a+bi)(c+di)) &= v((ac-bd)+(bc+ad)i) = \\ &= a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 = (a^2+b^2)(c^2+d^2) = v(a+bi)v(c+di) \end{aligned}$$

By Proposition 10.4, the units in $\mathbb{Z}[i]$ are exactly those elements $a+bi$ such that $v(a+bi) = a^2+b^2 = 1 = v(1)$, namely ± 1 and $\pm i$. This, together with the multiplicativity of v , allows us to prove the following fact:

Proposition 10.60 *If $r = a + bi \in \mathbb{Z}[i]$ and $v(r)$ is prime in $\mathbb{Z}$, then r is also prime in $\mathbb{Z}[i]$.*

Proof Suppose that $s|r$ in $\mathbb{Z}[i]$. Then $v(s)|v(r)$ in $\mathbb{Z}$. But $v(r)$ is prime in $\mathbb{Z}$, so $v(s)$ must equal either 1 or p . If $v(s) = 1$ then s is a unit in $\mathbb{Z}[i]$ by Proposition 10.4 and the above discussion. Otherwise, if $v(s) = p$ then $s \sim r$. In either case, p must be prime. $\square$

The ring $\mathbb{Z}[i]$ has a copy of $\mathbb{Z}$ embedded in it as a unital subring, so it's natural to ask what happens to the primes in $\mathbb{Z}$ when we consider them as elements of $\mathbb{Z}[i]$. Are they still prime in $\mathbb{Z}[i]$ or not? The answer is: sometimes, but not always. For example 2 is prime in $\mathbb{Z}$, but we can factorise it as $2 = (1+i)(1-i)$ in $\mathbb{Z}[i]$. Similarly, 5 and 13 are prime in $\mathbb{Z}$, but factorise as $5 = (1+2i)(1-2i)$ and $13 = (2+3i)(2-3i)$ in $\mathbb{Z}[i]$. On the other hand, 3, 7 and 11 are prime in both $\mathbb{Z}$ and $\mathbb{Z}[i]$. There are a couple of patterns here: one is that the prime integers that fail to be prime in $\mathbb{Z}[i]$ seem to factorise as a product of a Gaussian integer q with its conjugate $\bar{q}$; the other is that (apart from 2, which is often a special case) they are of the form $4n+1$.

We'll leave the second of these observations for a little while, but the first leads to the following fact.

Proposition 10.61 *Suppose $p \in \mathbb{Z}$ is prime in $\mathbb{Z}$. Then either p is also prime in $\mathbb{Z}[i]$, or it factorises as $p = q\bar{q}$, where q is prime in $\mathbb{Z}[i]$.*

Proof If $p \in \mathbb{Z}$ is not prime in $\mathbb{Z}[i]$, then there must exist two Gaussian integers $q, r \in \mathbb{Z}[i]$ such that $p = qr$, with at least one of q and r prime in $\mathbb{Z}[i]$, and neither of them are units.

Suppose that q is a Gaussian prime. Then by the multiplicativity of v ,

$$p^2 = v(p) = v(qr) = v(q)v(r).$$

Then $v(q) = v(r) = p$, so by Proposition 10.60 it follows that both q and r must be prime in $\mathbb{Z}[i]$. Also, $v(q) = q\bar{q} = p = qr$, so it must be the case that $r = \bar{q}$. $\square$

Now let's look at what happens to $v(q)$ when q is prime in $\mathbb{Z}[i]$. We know by Proposition 10.60 that if $v(q)$ is prime in $\mathbb{Z}$, then q must be prime in $\mathbb{Z}[i]$, but the converse doesn't necessarily hold: some Gaussian primes q might yield non-prime values of $v(q)$. Applying v to some small Gaussian primes, we see the following:

$$\begin{aligned} v(1+i) &= 2, & v(1+2i) &= 5, & v(3) &= 9, & v(2+3i) &= 13, \\ v(1+4i) &= 17, & v(2+5i) &= 29, & v(7) &= 49, & v(11) &= 121. \end{aligned}$$

Again, various patterns are starting to form: $v(q)$ is either a prime integer, or the square of a prime integer. Also, in the cases when $v(q)$ is the square of a prime integer, q is an integer of the form $4n-1$, and in the cases where q is a Gaussian prime with nonzero imaginary part, $v(q)$ is of the form $4n+1$. Except, that is, for the special case where $v(q) = 2$, but 2 is often a special case when discussing prime numbers.

The first of these is relatively straightforward to prove:

Proposition 10.62 *Let $q \in \mathbb{Z}[i]$ be a Gaussian prime. Then $v(q)$ is either a prime integer or the square of a prime integer.*

Proof Let $v(q) = n$, and factorise $n = p_1 \dots p_k$ as a product of primes in $\mathbb{Z}$. Recall that $v(q) = q\bar{q}$, so we have $q\bar{q} = p_1 \dots p_k$. Then $q|p_1 \dots p_k$, and since q is prime in $\mathbb{Z}[i]$, it must be the case that $q|p_i$ for some i . Then $v(q)|v(p_i)$, and $v(p_i) = p_i^2$, so $v(q)|p_i^2$. So either $v(q) = p_i$ or $v(q) = p_i^2$, as claimed. $\square$

We are now in a position to classify Gaussian primes:

Proposition 10.63 *Let $q \in \mathbb{Z}[i]$ be a Gaussian prime. Then q is one of the following three types:*

- (i) $q \sim 1 \pm i$, with $v(q) = 2$;
- (ii) $q = a + bi$, with $v(q) = a^2 + b^2 \equiv 1 \pmod{4}$; or
- (iii) $q = a$, where a is prime in $\mathbb{Z}$ and $a \equiv 3 \pmod{4}$.

Proof By Proposition 10.62, $v(q)$ is either a prime integer p , or its square p^2 . If $v(q) = p$ then by Proposition 10.61 we have $v(q) = p = q\bar{q}$.

If, on the other hand, $v(q) = p^2$ then $q|p^2$ in $\mathbb{Z}[i]$, and since q is prime in $\mathbb{Z}[i]$ we must also have $q|p$ in $\mathbb{Z}[i]$, so there exists some $r \in \mathbb{Z}[i]$ such that $qr = p$. Then $v(r) = v(p)/v(q) = p^2/p^2 = 1$, so r must be a unit in $\mathbb{Z}[i]$, and thus $q = \pm p$ or $\pm ip$. That is, if $v(q)$ is the square of a prime integer p , then q must be an associate in $\mathbb{Z}[i]$ of p .

We now look at the modulo-4 congruence of this prime p . Suppose that $q = a + bi$. Then $v(q) = q\bar{q} = a^2 + b^2$, and with the congruences

$$\begin{aligned} 0^2 &\equiv 0 \pmod{4}, & 1^2 &\equiv 1 \pmod{4}, \\ 2^2 &\equiv 0 \pmod{4}, & 3^2 &\equiv 1 \pmod{4}, \end{aligned}$$

it's clear that we can't have $a^2 + b^2 \equiv 3 \pmod{4}$. So if $p \equiv 3 \pmod{4}$ then it can't be factorised as $p = q\bar{q}$ in $\mathbb{Z}[i]$, and therefore p must be prime in $\mathbb{Z}[i]$ too.

As noted earlier, $2 = (1+i)(1-i)$, and both $(1+i)$ and $(1-i)$ are a conjugate pair of Gaussian primes. Also, $(1+i) = -i(1-i)$ and $(1-i) = i(1+i)$ so they are associates.

By Corollary 10.10 the multiplicative group $U(\mathbb{F}_p) = \mathbb{Z}_p \setminus \{0\}$ is cyclic of order $p-1$, and is thus isomorphic to $\mathbb{Z}_{p-1}$. Now suppose that $p \equiv 1 \pmod{4}$; this means that $4|p-1$, so this group $\mathbb{Z}_{p-1}$ has an element a of order 4.

The polynomial $x^2 - 1$ has at most two roots in $\mathbb{Z}_p$, namely 1 and $-1 \equiv p-1 \pmod{p}$. Then -1 is the only nonzero element of order 2 in $\mathbb{Z}_p$. Since a has order 4, it follows that a^2 has order 2, and

hence $a^2 = -1$ in $\mathbb{Z}_p$. This means that $a^2 \equiv p-1 \pmod p$ in $\mathbb{Z}$, and so $a^2 + 1 \equiv 0 \pmod p$. Therefore there exists $k \in \mathbb{Z}$ such that $a^2 + 1 = kp$, and hence $(a+i)(a-i) = kp$ in $\mathbb{Z}[i]$, which means that $p|(a+i)(a-i)$ in $\mathbb{Z}[i]$. If p were a prime in $\mathbb{Z}[i]$ then either $p|(a+i)$ or $p|(a-i)$. But neither of these can happen, because multiples of p in $\mathbb{Z}[i]$ are of the form $p(r+si) = pr + psi$ for some $r, s \in \mathbb{Z}$. Hence p can't be prime in $\mathbb{Z}[i]$ if $p \equiv 1 \pmod 4$.

This completes the proof: $v(q)$ is either a prime p , in which case $p = q\bar{q}$ and q is either $(1 \pm i)$ and $p = 2$, or $p \equiv 1 \pmod 4$; or $v(q) = p^2$, in which case $p \equiv 3 \pmod 4$ and $q \sim p$. $\square$

We can use this to prove a result originally proved by Pierre de Fermat (1601–1665).

Corollary 10.64 *Let p be a prime integer such that either $p = 2$ or $p \equiv 1 \pmod 4$. Then p can be expressed uniquely as a sum of positive integer squares.*

Proof If $p = 2$ or $p \equiv 1 \pmod 4$ then $p = q\bar{q} = a^2 + b^2$ for some Gaussian prime $q = a + bi \in \mathbb{Z}[i]$, by Proposition 10.63.

To show uniqueness, suppose that $p = a^2 + b^2 = c^2 + d^2$ for some integers $a, b, c, d > 0$. Then $p = (a + bi)(a - bi) = (c + di)(c - di)$ are two prime decompositions of p in $\mathbb{Z}[i]$. But $\mathbb{Z}[i]$ is a UFD, so these factorisations are the same up to permutation and multiplication by a unit $\pm 1, \pm i$. So either $(a + bi) = \pm(c + di)$ or $(a + bi) = \pm i(c + di) = (\mp d \pm ci)$. Hence either $a = c$ and $b = d$ or $a = d$ and $b = c$. $\square$

The ring $\mathbb{Z}[i]$ of Gaussian integers is one of a class of integral domains called **quadratic integer rings**. These are of the form $\mathbb{Z}[\alpha]$ where α is a root of a **monic** quadratic polynomial (that is, a quadratic polynomial whose top coefficient is 1).

Definition 10.65 Let $f = x^2 + bx + c \in \mathbb{Z}[x]$ be a quadratic polynomial with integer coefficients, and let α be a root of this polynomial. Then

$$\mathbb{Z}[\alpha] = \{s + t\alpha : s, t \in \mathbb{Z}\}$$

is the **quadratic integer ring** determined by f and α . We say $\mathbb{Z}[\alpha]$ is **real** if $\alpha \in \mathbb{R}$, and **imaginary** otherwise.

We've seen several examples of quadratic integer rings already: $\mathbb{Z}[i]$ is determined by the polynomial $x^2 + 1$ and the root $x = i$; we obtain an isomorphic ring if we use the other root $x = -i$, the isomorphism being the usual conjugation map $z \mapsto \bar{z}$.

The real quadratic integer ring $\mathbb{Z}[\sqrt{2}]$ is determined by the polynomial $x^2 - 2$ and the root $x = \sqrt{2}$. Again, this is isomorphic to the ring $\mathbb{Z}[-\sqrt{2}]$ by the map $f: (a+b\sqrt{2}) \mapsto (a-b\sqrt{2})$.

Now let $f = x^2 + x + 1$ and let $\omega = \frac{1}{2} + \frac{\sqrt{3}}{2}i$ be one of its roots. We then obtain the ring $\mathbb{Z}[\omega]$ of Eisenstein integers. Again, $\mathbb{Z}[\omega] \cong \mathbb{Z}[-\omega]$.

What we want to do now is to generalise some of the constructions we used for the Gaussian integers, in particular the concept of a conjugate, and the Euclidean valuation or norm $v: \mathbb{Z}[i] \setminus \{0\} \rightarrow \mathbb{N} \cup \{0\}$.

The first step is to generalise the notion of a complex conjugate to an arbitrary quadratic integer ring $\mathbb{Z}[\alpha]$. The key is to look at the conjugation isomorphism $\mathbb{Z}[i] \rightarrow \mathbb{Z}[-i]$, and the isomorphism $\mathbb{Z}[\sqrt{2}] \rightarrow \mathbb{Z}[-\sqrt{2}]$. In both cases, we swapped the chosen root α of the defining polynomial f with the other root β . This approach works in general:

Definition 10.66 Suppose that α is a root of $f = x^2 + bx + c$, and $\mathbb{Z}[\alpha]$ is the corresponding quadratic integer ring. Let β be the other root of f ; by the usual quadratic formula, $\beta = -b - \alpha$. Then the **conjugate** of an element $r = (s + t\alpha) \in \mathbb{Z}[\alpha]$ is defined to be

$$\bar{r} = s + t\beta = s - tb - t\alpha.$$

For $\mathbb{Z}[i]$ we used the valuation function

$$v(a+bi) = (a+bi)\overline{(a+bi)} = (a+bi)(a-bi) = a^2 + b^2.$$

This valuation has all the necessary properties to ensure $\mathbb{Z}[i]$ is a Euclidean domain, and was also very useful when we studied the Gaussian primes. And now that we have a general notion of a conjugate in a given quadratic integer ring, the following definition seems the natural next step.

Definition 10.67 Suppose that α, β are the roots of a monic quadratic polynomial $f = x^2 + bx + c \in \mathbb{Z}[x]$, and that $\mathbb{Z}[\alpha]$ is the corresponding quadratic integer ring. Let $r = s + t\alpha$ be an arbitrary element of $\mathbb{Z}[\alpha]$. Then we define the **norm** of r to be

$$N(r) = r\bar{r} = s^2 - bst + ct^2.$$

This norm will sometimes satisfy the conditions for a Euclidean valuation, either on its own or by defining $v(r) = |N(r)|$, but sometimes it won't. Rings for which the absolute value $|N(r)|$ of the norm N happens to be a Euclidean valuation are said to be **norm Euclidean**. However, not every Euclidean domain is norm Euclidean in this sense: just because the absolute value of the norm N isn't a Euclidean valuation doesn't mean that no such valuation exists.¹⁷

Although the norm of a quadratic integer ring needn't give a Euclidean valuation, it does have some useful properties, in particular the following.

¹⁷ For example, $\mathbb{Z}[\sqrt{14}]$ is not norm Euclidean, but was shown in 2004 to nevertheless be a Euclidean domain anyway; however, the techniques used are beyond the scope of this book.¹⁸

¹⁸ M Harper, $\mathbb{Z}[\sqrt{14}]$ is Euclidean, Canadian Journal of Mathematics 36 (2004) 55–70

Proposition 10.68 *Let $\mathbb{Z}[\alpha]$ be a quadratic integer ring determined by a root α of a monic polynomial $f = x^2 + bx + c \in \mathbb{Z}[x]$, and let $N(r) = r\bar{r}$ be the corresponding norm. Then $N(rs) = N(r)N(s)$ for all $r, s \in \mathbb{Z}[\alpha]$, and $N(r) = \pm 1$ if and only if r is a unit of $\mathbb{Z}[\alpha]$.*

Proof The multiplicative property can be verified by a straightforward but tedious calculation, which we will therefore skip here, but which the reader is invited to do for themselves.

If $rs = 1$ in $\mathbb{Z}[\alpha]$, then taking norms of both sides we have

$$N(r)N(s) = N(rs) = N(1) = 1.$$

So, $N(r) = N(s) = \pm 1$. Conversely, if $N(r) = \pm 1$ then $r\bar{r} = \pm 1$, which means that r must be invertible with $r^{-1} = \pm\bar{r}$. $\square$

What we want to do is to find an example of a quadratic integer ring that is a PID but isn't a Euclidean domain. In order to do this, we need to introduce a little more technology: A necessary condition for something to be a Euclidean domain, and a sufficient condition for something to be a PID. Then if we can find something that satisfies one and not the other, we'll have the example we're looking for.

Taking the first of these, we'll have a look at some examples of quadratic integer rings that are Euclidean.

Our motivating example of a Euclidean domain, as well as a lot of the other algebraic structures we've investigated so far, is the ring $\mathbb{Z}$ of integers. We can partition this into even and odd numbers: every element of $\mathbb{Z}$ is either even or odd, having the form $2n$ or $2n+1$. One way of looking at this is to say, as we did in the last chapter, that $\mathbb{Z}/2\mathbb{Z} \cong \mathbb{Z}_2$. But now we're interested in the existence or nonexistence of a Euclidean valuation $v: \mathbb{Z} \rightarrow \mathbb{N} \cup \{0\}$.

Both the multiplicative norm N of a quadratic integer ring $\mathbb{Z}[\alpha]$ and the valuation defined on a Euclidean domain (and also the Dedekind–Hasse norms we'll meet in a little while) can be thought of as distance functions in the domain in question. So another way of looking at the even/odd partition of $\mathbb{Z}$ is to say that every element of $\mathbb{Z}$ is at most distance 1 from an even integer, as depicted in Figure 10.2. The same holds for multiples of 3: every integer is of the form $3n-1$, $3n$ or $3n+1$, as shown in Figure 10.3. However, this doesn't work for multiples of 4: there exist infinitely many integers of the form $4n+2$, which are distance 2 from the nearest multiple of 4, as shown in Figure 10.4.

We can do this with the Gaussian integers too: if we consider the ideal $\langle 1+i \rangle$ generated by the element $1+i$, we get the lattice depicted in Figure 10.5. Every element of $\mathbb{Z}[i]$ is either a point in this lattice (that is, an element of $\langle 1+i \rangle$) or distance 1 from it using the valuation

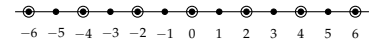


Figure 10.2: Even and odd integers

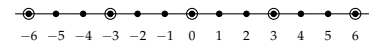


Figure 10.3: Multiples of 3

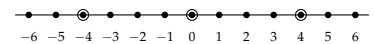


Figure 10.4: Multiples of 4

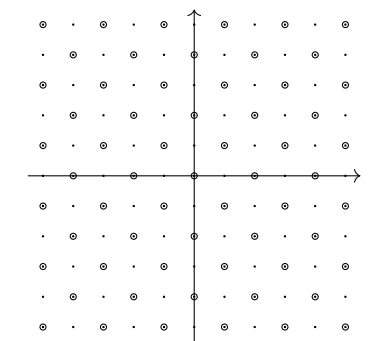


Figure 10.5: The ideal $\langle 1+i \rangle$ in $\mathbb{Z}[i]$

$$v(a+bi) = a^2 + b^2.$$

Alternatively, every element in $\mathbb{Z}[i]$ can be expressed in the form $z+a$ where $a \in \langle 1+i \rangle$ and $z \in \{0, \pm 1, \pm i\}$. Similarly, every element of $\mathbb{Z}$ can be written as $m+a$ where $a \in \langle 2 \rangle$ and $m = 0, \pm 1$; or in the form $m+b$ where $b \in \langle 3 \rangle$ and $m = 0, \pm 1$.

This, we'll see in a moment, is a property shared by every Euclidean domain R . Given a Euclidean valuation defined on R , we can find a nonzero element $r \in R$ such that the cosets of $R/\langle r \rangle$ are of the form $a+\langle r \rangle$ where a is either 0 or a unit of R . Time for some terminology:

Definition 10.69 Let R be an integral domain, and let $r \in R$ be such that every coset in $R/\langle r \rangle$ can be represented as $a+\langle r \rangle$ where either $a = 0$, or a is a unit of R . Then r is a **universal side divisor** of R .

This is the key concept we need: it turns out that a Euclidean domain must have at least one universal side divisor.

Proposition 10.70 Let R be a Euclidean domain, and let v be a Euclidean valuation defined on R . Then let $a \in R$ be a nonunit element of R for which $v(a)$ is minimal over all nonunits in R . Then a is a universal side divisor.

Proof Let $x \in R$ be any element of R . Then since R is a Euclidean domain, we can find $q, r \in R$ such that $x = qa + r$ and either $r = 0$ or $v(r) < v(a)$.

If $r \neq 0$ then $v(r) < v(a)$ and by Proposition 10.4 r must be a unit. Since $a|(x-r)$, the quotient $R/\langle a \rangle$ must be represented by 0 and the units in R . $\square$

Our current goal is to show that $\mathbb{Z}[\alpha]$, where $\alpha = \frac{1}{2}(1+\sqrt{-19})$, is a PID but not a Euclidean domain. If we can show it doesn't have any universal side divisors, then by Proposition 10.70 it can't be a Euclidean domain.

Proposition 10.71 Let $\mathbb{Z}[\alpha]$ be the quadratic integer ring determined by the polynomial $x^2 - x + 5$ and the root $\alpha = \frac{1}{2}(1+\sqrt{-19})$. Then $\mathbb{Z}[\alpha]$ has no universal side divisors and hence is not a Euclidean domain.

Proof First we want to find the units in $\mathbb{Z}[\alpha]$, and for this we use the norm $N(a + b\alpha) = a^2 + ab + 5b^2$. This is **positive definite**; that is, $N(a + b\alpha) \geq 0$ for all $a, b \in \mathbb{Z}[\alpha]$, and $N(a + b\alpha) = 0$ only when $a = b = 0$. We can show this by completing the square, which yields the two expressions

$$\left(a + \frac{1}{2}b\right)^2 + \frac{19}{4}b^2 \quad \text{and} \quad \frac{19}{20}a^2 + \left(\frac{1}{2\sqrt{5}}a + \sqrt{5}b\right)^2$$

for $N(a + b\alpha)$. The first of these is neater than the second, but both are clearly positive definite. If $b \neq 0$ then

$$N(a + b\alpha) = \left(a + \frac{1}{2}b\right)^2 + \frac{19}{4}b^2 \geq \frac{19}{4}b^2 \geq \frac{19}{4} > 4.$$

By Proposition 10.68 the units are exactly those elements $r = (a + b\alpha) \in \mathbb{Z}[\alpha]$ for which $N(r) = 1$, so these must have $b = 0$, and hence we require $N(r) = N(a) = a^2 = 1$. Thus $a = \pm 1$ and so the units in $\mathbb{Z}[\alpha]$ are ± 1 .

Now suppose that $\mathbb{Z}[\alpha]$ has a universal side divisor d . Then for any $r \in \mathbb{Z}[\alpha]$ there exists some $u \in \{0, \pm 1\}$ such that $r \equiv u \pmod{d}$ in $\mathbb{Z}[\alpha]$. This means that $r = kd + u$ for some $k \in \mathbb{Z}[\alpha]$, and hence $r - u = kd$. Thus $d \mid (r - u)$.

Suppose $r = 2$. Then $d \mid (2 - u)$ for some $u \in \{0, \pm 1\}$. If $u = 1$ then $d \mid 1$, so d is a unit. But Definition 10.69 specifically rules this out: a universal side divisor may not be a unit. If $u = 0$ then $d \mid 2$, and if $u = -1$ then $d \mid 3$, so the only possibilities open to us are that $d \in \{\pm 2, \pm 3\}$. We will now rule out each of these four possibilities as well.

Consider another element $r \in \mathbb{Z}[\alpha]$: take $r = \alpha = \frac{1}{2}(1 + \sqrt{-19})$ itself. Again, $d \mid (r - u)$ for some $u \in \{0, \pm 1\}$; that is, either $d \mid (r - 1)$, or $d \mid r$, or $d \mid (r + 1)$.

Suppose that $d \mid (r + 1)$. Then there exists some $s \in \mathbb{Z}[\alpha]$ such that $sd = r + 1 = \frac{3}{2} + \frac{1}{2}\sqrt{-19}$. Applying the norm N to this, we get

$$N(s)N(d) = N(sd) = N(1 + \alpha) = 7.$$

This can't happen, because either $N(d) = 4$ or 9 , and neither of these are factors of 7 .

If $d \mid r$ then $sd = r = \alpha$, and $N(s)N(d) = N(sd) = N(\alpha) = 5$; this is also prime, so there is no suitable s that can be multiplied by any $d \in \{\pm 2, \pm 3\}$ to give α .

Finally, if $d \mid (r - 1)$ then there would exist some $s \in \mathbb{Z}[\alpha]$ such that $sd = \alpha - 1 = -\frac{1}{2} + \frac{1}{2}\sqrt{-19}$. Using the norm again, this would require $N(s)N(d) = N(sd) = N(\alpha - 1) = 7$, which can't happen for the same reasons as in the previous two cases.

Therefore there is no universal side divisor $d \in \mathbb{Z}[\alpha]$, and hence $\mathbb{Z}[\alpha]$ can't be a Euclidean domain. $\square$

Next we need to show that $\mathbb{Z}[\alpha]$ is a PID, and to do that we have to introduce another bit of algebraic machinery. Recall that a Euclidean domain is an integral domain for which we can define a Euclidean norm or Euclidean valuation. Is there a slightly more general class of norms whose existence ensures that a given integral domain is a PID? The answer is yes:

Definition 10.72 Let R be an integral domain, and let $N: R \rightarrow \mathbb{Z}$ be a function defined on R . If N is positive definite and if, for every $a, b \in R$, either

(i) $a \mid b$, or

(ii) there exists a nonzero element $c \in \langle a, b \rangle$ such that

$$0 < N(c) < N(b),$$

then N is a **Dedekind–Hasse norm**.

The first of these conditions is equivalent to requiring that $b \in \langle a \rangle$, and the second is equivalent to the existence of elements $s, t \in R$ such that

$$0 < N(sa + tb) < N(b).$$

Every Euclidean valuation is a Dedekind–Hasse norm: the Euclidean condition is the same as saying we can always set $s = 1$, so that $0 < N(a + tb) < N(b)$.

This new type of norm is exactly what we need:

Proposition 10.73 *An integral domain R is a PID if and only if there exists a Dedekind–Hasse norm defined on R .*

Proof Let I be an ideal in R and choose $b \in I$ such that $N(b)$ is less than the norm of any other element in I . Then let $a \in I$ be some arbitrary nonzero element of I . Since I is an ideal of R , then for any $s, t \in R$ it must be the case that $(sa - tb) \in I$. Then either $(sa - tb) = 0$ or $N(sa - tb) \geq N(b)$ for any $s, t \in R$. Hence, by Definition 10.72, since $N(sa - tb) \not< N(b)$ for any $s, t \in R$, it must be the case that $b|a$, and thus $a \in \langle b \rangle$. Thus $I = \langle b \rangle$ is principal, and hence R must be a PID.

Conversely, suppose that R is a PID. We will construct a Dedekind–Hasse norm on R . Let $N(0) = 0$, and let $N(u) = 1$ for any unit $u \in R$. Since every PID is a UFD by Corollary 10.43, any nonzero nonunit r in R can be factorised into a finite product of irreducible elements. Any such factorisation is unique up to permutation and multiplication by units, and thus will have the same number n_r of irreducible elements. Define $N(r) = 2^{n_r}$. Then N is positive definite.

It also satisfies the multiplicativity condition, since

$$N(rs) = 2^{n_{rs}} = 2^{n_r + n_s} = 2^{n_r} 2^{n_s} = N(r)N(s).$$

We claim that this is a Dedekind–Hasse norm on R . Suppose $a, b \in R$ with $b \neq 0$. Then $I = \langle a, b \rangle = \{as + bt : s, t \in R\}$ is an ideal of R , and since R is a PID, there exists some nonzero element $d \in R$ such that $I = \langle d \rangle$. This element $d = as + bt$ for some $s, t \in R$.

Then $b = rd \in I = \langle d \rangle$ for some nonzero $r \in R$, so

$$N(b) = N(rd) = N(r)N(d) > N(d) = N(as + bt) > 0.$$

Thus N is a Dedekind–Hasse norm for R . □

So, if we can show now that $\mathbb{Z}[\alpha]$ does admit a Dedekind–Hasse norm then we’re done: we’ve found a PID that isn’t Euclidean. The most

straightforward approach is to confirm that the norm $N(a + b\alpha) = a^2 + ab + 5b^2$ satisfies either of the Dedekind–Hasse criteria.

Proposition 10.74 *Let $\mathbb{Z}[\alpha]$ be the quadratic integer ring determined by the polynomial $f = x^2 - x + 5 \in \mathbb{Z}[x]$ and the root $\alpha = \frac{1}{2}(1 + \sqrt{-19})$. Then $N(a + b\alpha) = a^2 + ab + 5b^2$ is a Dedekind–Hasse norm, and hence $\mathbb{Z}[\alpha]$ is a PID.*

The following proof is based on the version given by Wilson.¹⁹

Proof Suppose $u, v \in \mathbb{Z}[\alpha]$ with $v \neq 0$ and $v \nmid u$, and $N(u) \geq N(v)$. We want to find elements $s, t \in \mathbb{Z}[\alpha]$ such that $0 < N(su + tv) < N(v)$. Since N satisfies the multiplicative condition, this is the same as finding $s, t \in \mathbb{Z}[\alpha]$ such that

$$0 < N\left(\frac{u}{v}s - t\right) < N(1) = 1.$$

In general, the quotient $\frac{u}{v}$ doesn't lie in $\mathbb{Z}[\alpha]$; indeed, unless v is a unit or $v|u$, this doesn't even make sense in $\mathbb{Z}[\alpha]$. Here, we've specifically chosen u and v so that this is the case, so we must instead consider $\frac{u}{v}$ as an element of $Q(\mathbb{Z}[\alpha])$, the field of fractions of $\mathbb{Z}[\alpha]$. This is exactly

$$\mathbb{Q}[\alpha] = \{a + b\alpha : a, b \in \mathbb{Q}\} = \{a + b\sqrt{-19} : a, b \in \mathbb{Q}\} = \mathbb{Q}[\sqrt{-19}].$$

The norm N extends happily to $\mathbb{Q}[\alpha]$, but when we switch to $\mathbb{Q}[\sqrt{-19}]$ it takes a slightly different form:

$$N(a + b\sqrt{-19}) = a^2 + 19b^2$$

Suppose, then, that $\frac{u}{v} = \frac{1}{c}(a + b\sqrt{-19}) \in \mathbb{Q}[\sqrt{-19}]$, where $a, b, c \in \mathbb{Z}$ are coprime integers and $c > 1$. (This latter condition must be true, because we've already decided that $v \nmid u$.) Since a, b, c are coprime, and hence $\gcd(a, b, c) = 1$, there must exist integers l, m, n such that $la + mb + nc = 1$.

Furthermore, there must exist integers $q, r \in \mathbb{Z}$ such that $am - 19bl = cq + r$ with $|r| \leq \frac{1}{2}c$. We have four cases to consider.

Case 1 ($c \geq 5$) Set $s = m + l\sqrt{-19}$ and $t = q - n\sqrt{-19}$. Then

$$\begin{aligned} 0 < N\left(\frac{u}{v}s - t\right) &= N\left(\frac{1}{c}(a + b\sqrt{-19})(m + l\sqrt{-19}) - (q - n\sqrt{-19})\right) \\ &= N\left(\frac{1}{c}((am - 19bl) + (bm + al)\sqrt{-19}) - q + n\sqrt{-19}\right) \\ &= N\left(\frac{1}{c}(am - 19bl - cq) + (al + bm + cn)\sqrt{-19}\right) \\ &= N\left(\frac{1}{c}(r + \sqrt{-19})\right) \\ &= \frac{r^2 + 19}{c^2}. \end{aligned}$$

If $c > 5$ then $\frac{r^2 + 19}{c^2} \leq \frac{1}{4} + \frac{19}{c^2} \leq \frac{7}{9} < 1$, so N is a Dedekind–Hasse norm.

If $c = 5$ then $|r| \leq 2$ and so $\frac{r^2 + 19}{c^2} \leq \frac{23}{25} < 1$, and again N is a Dedekind–Hasse norm.

¹⁹ JC Wilson, *A principal ideal ring that is not a Euclidean ring*, Mathematics Magazine 46 (1973) 34–38.

Case 2 ($c = 2$) One of a and b is even, and the other must be odd, otherwise we would have $\frac{u}{v} \in \mathbb{Z}[\alpha]$. Let

$$s = 1 \quad \text{and} \quad t = \frac{1}{2}(a-1) + \frac{1}{2}b\sqrt{-19} = \frac{1}{2}(a-b-1) + \frac{1}{2}b\alpha \in \mathbb{Z}[\alpha].$$

Then

$$N\left(\frac{u}{v}s - t\right) = N\left(\frac{1}{2}(a+b\sqrt{-19}) - \frac{1}{2}(a-1) - \frac{1}{2}b\sqrt{-19}\right) = N\left(\frac{1}{2}\right) < 1.$$

Case 3 ($c = 3$) If $c = 3$ then 3 doesn't divide both a and b , so $a^2 + 19b^2$ isn't divisible by 3 either. Choose $q, r \in \mathbb{Z}$ such that $a^2 + 19b^2 = 3q + r$, with $0 < r < 3$; that is, $r = 1$ or 2 . Set

$$s = a - b\sqrt{-19} \quad \text{and} \quad t = q \in \mathbb{Z}[\alpha].$$

Then

$$\begin{aligned} N\left(\frac{u}{v}s - t\right) &= N\left(\frac{1}{3}(a+b\sqrt{-19})(a-b\sqrt{-19}) - q\right) \\ &= N\left(\frac{1}{3}(a^2 + 19b^2) - q\right) \\ &= N\left(\frac{1}{3}(3q + r) - q\right) \\ &= N\left(\frac{1}{3}r\right) < 1. \end{aligned}$$

Case 4 ($c = 4$) Finally, suppose that $c = 4$. Then at most one of a and b can be even, otherwise this reduces to Case 2. If only one is odd then $a^2 + 19b^2$ is odd, so we can find $q, r \in \mathbb{Z}$ such that $a^2 + 19b^2 = 4q + r$, with $r = 1$ or 3 . As with Case 3, set

$$s = a - b\sqrt{-19} \quad \text{and} \quad t = q \in \mathbb{Z}[\alpha].$$

Then

$$N\left(\frac{u}{v}s - t\right) = N\left(\frac{1}{4}(a+b\sqrt{-19})(a-b\sqrt{-19}) - q\right) = N\left(\frac{1}{4}r\right) < 1.$$

If both a and b are odd, then $a^2 + 19b^2 \not\equiv 0 \pmod{8}$, so choose $q, r \in \mathbb{Z}$ such that $a^2 + 19b^2 = 8q + r$, with $0 < r < 8$. Set

$$s = \frac{1}{2}(a - b\sqrt{-19}) \quad \text{and} \quad t = q.$$

Then

$$N\left(\frac{u}{v}s - t\right) = N\left(\frac{1}{8}(a+b\sqrt{-19})(a-b\sqrt{-19}) - q\right) = N\left(\frac{1}{8}r\right) < 1.$$

In each case, then, N is a Dedekind–Hasse norm, and so by Proposition 10.73 $\mathbb{Z}[\alpha]$ is a PID. $\square$

We have now proved the existence of at least one PID that isn't a Euclidean domain, and thereby completed the construction of the hierarchy

$$\{\text{Euclidean domains}\} \subset \{\text{PIDs}\} \subset \{\text{UFDs}\} \subset \{\text{integral domains}\}.$$

Each inclusion in this hierarchy is strict.

Summary

References and further reading

Exercises

- 10.1 List all the primitive polynomials of degree 3 in $\mathbb{Z}_5[x]$.
- 10.2 Show that the ring $\mathbb{Z}[\omega]$ of Eisenstein integers is a Euclidean domain via the valuation $v(a + b\omega) = a^2 - ab + b^2$.
- 10.3 Let $\mathbb{F}$ be a field, and suppose that $f \in \mathbb{F}[x]$ is irreducible. Show that $\mathbb{F}[x]/\langle f \rangle$ is a field.

11 Polynomials

I'm very well acquainted, too,
with matters mathematical,
I understand equations,
both the simple and quadratical,
About binomial theorem
I'm teeming with a lot o' news,
With many cheerful facts
about the square of the hypotenuse.

— WS Gilbert (1836–1911) and
Arthur Sullivan (1842–1900),
The Major General's Song from
The Pirates of Penzance (1879)

THE PROBLEM of finding solutions to polynomial equations is an old one with an interesting and occasionally scandalous or tragic history. In this chapter we will study the factorisation problem for polynomials over a field, by introducing the idea of a **field extension**, which will lead us to investigate field automorphisms and construct the **Galois group** of a polynomial. This group, named after and originally introduced by the young French mathematician Évariste Galois (1811–1832), enables us to determine whether the polynomial has an algebraic solution or **solution by radicals**.

11.1 Irreducible polynomials

In this chapter we will mainly be concerned with polynomials over fields; that is, elements of $F[x]$ where F is a field. We know from Theorem 10.5 that $F[x]$ is a Euclidean domain, and the Remainder Theorem¹ tells us that finding the roots of such a polynomial is equivalent to factorising it into linear factors. And since $F[x]$ is a Euclidean domain, it must also be a UFD, so any such irreducible factorisation is unique up to permutation and multiplication by units. So to properly study the solution of polynomial equations, we need to understand the irreducible elements in $F[x]$.

Gauss' Lemma, as stated in Proposition 10.51, says that the products of primitive polynomials (those whose coefficients are coprime) are also primitive. Proposition 10.55 tells us that if R is a UFD and $Q = Q(R)$ is its field of quotients, then a primitive polynomial $f \in R[x]$ is irreducible in $R[x]$ if and only if it is irreducible in $Q[x]$. (The special case when $R = \mathbb{Z}$ and $Q = \mathbb{Q}$ is also sometimes called Gauss' Lemma.)

So, by Corollary 10.56, an irreducible element in $R[x]$ is either irreducible in R or a primitive polynomial in $R[x]$ that is also irreducible in $Q[x]$. Another important result on irreducible polynomials is due

Matter, though divisible in an extreme degree, is nevertheless not infinitely divisible. That is, there must be some point beyond which we cannot go in the division of matter. The existence of these ultimate particles of matter can scarcely be doubted, though they are probably much too small ever to be exhibited by microscopic improvements. I have chosen the word *atom* to signify these ultimate particles in preference to *particle*, *molecule*, or any other diminutive term, because I conceive it is much more expressive; it includes in itself the notion of *indivisible*, which the other terms do not.

— John Dalton (1766–1844),
lecture at the Royal Institution,
30 January 1810

¹ Theorem 10.7, page 382.



Wikimedia Commons / Unknown artist

Ferdinand Gotthold Max Eisenstein (1823–1852) was born in Berlin and displayed a strong aptitude in both mathematics and music from an early age. At 17, having exhausted the school mathematical curriculum, he began attending lectures at the University of Berlin, and formally enrolled in 1843.

After graduation, he met the geographer Alexander von Humboldt (1769–1859), who managed to obtain grants to support his career. He also visited Carl Friedrich Gauss (1777–1855), who was impressed: the mathematical historian Moritz Cantor (1829–1920), one of Gauss' last students, later wrote that he "once remarked in conversation that there had been only three epoch-making mathematicians: Archimedes, Newton and Eisenstein."

Over the rest of his life, he made many contributions to the study of quadratic and cubic forms, and to reciprocity laws for prime numbers. He was also an accomplished pianist and composer. Eisenstein's physical health was delicate throughout his life, due to meningitis he contracted during childhood (and which also killed his five siblings); also he suffered from depression, a legacy of an overly harsh school environment. In 1848 he found himself in the middle of a political demonstration and was arrested. Although released the next day, the incident contributed to a further decline in his health, and caused the loss of some of his funding. He died of tuberculosis in 1852, at the age of 29; Humboldt led the funeral procession.

to the German mathematician Gotthold Eisenstein (1823–1852).

Proposition 11.1 (Eisenstein's Criterion) *Suppose that*

$$f = a_n x^n + \cdots + a_1 x + a_0 \in \mathbb{Z}[x]$$

is a polynomial with integer coefficients, and that there is a prime p such that

- (i) $p \nmid a_n$,
- (ii) $p \mid a_{n-1}, \dots, a_0$, and
- (iii) $p^2 \nmid a_0$.

Then f is irreducible over $\mathbb{Q}$.

Proof If we can show that f is irreducible over $\mathbb{Z}$, then Gauss' Lemma ensures that it will also be irreducible over $\mathbb{Q}$.

Suppose, then, that f is not irreducible, and that there exists a prime p satisfying the hypotheses. Then $f = gh$ for some polynomials

$$g = b_l x^l + \cdots + b_1 x + b_0 \quad \text{and} \quad h = c_m x^m + \cdots + c_1 x + c_0$$

in $\mathbb{Z}[x]$. Since $a_0 = b_0 c_0$ and $p \mid a_0$, we know that either $p \mid b_0$ or $p \mid c_0$. But p can't divide both b_0 and c_0 , since if it did we would have $p^2 \mid a_0$, which the third hypothesis disallows. We can assume without loss of generality that $p \mid b_0$ and $p \nmid c_0$.

Furthermore, the leading coefficient $a_n = b_l c_m$ is not divisible by p , so $p \nmid b_l$ and $p \nmid c_m$.

Now let b_i be the first coefficient in g that isn't divisible by p , so that $p \mid b_0, \dots, b_{i-1}$. If $i < n$ then $p \mid a_i$ by the second hypothesis. Also, since

$$b_i c_0 = a_i - (b_{i-1} c_1 + \cdots + b_0 c_i),$$

then we must have $p \mid b_i c_0$. But this can't happen, because $p \nmid b_i$ and $p \nmid c_0$. So instead $i = n$, which means that $\deg(g) = n$ and $\deg(h) = 0$. Therefore h is a nonzero constant $c_0 \in \mathbb{Z}$, and so $f = c_0 g$.

If f is irreducible over $\mathbb{Z}$ then by Gauss' Lemma it is irreducible over $\mathbb{Q}$; if not then it has the form $f = c_0 g$ for some nonzero $c_0 \in \mathbb{Z}$, which is a unit in $\mathbb{Q}$, and therefore f is again irreducible over $\mathbb{Q}$. $\square$

Example 11.2 The polynomial $f = 2x^3 + 6x + 3$ is irreducible over $\mathbb{Q}$: setting $p = 3$ we have $p \nmid a_3$, while $p \mid a_1$ and $p \mid a_0$, but $p^2 \nmid a_0$.

Another useful technique for checking irreducibility is reduction modulo p , for some carefully chosen prime p . For any $n \in \mathbb{N}$ there is a canonical quotient homomorphism $q_n: \mathbb{Z} \rightarrow \mathbb{Z}_n$ mapping each integer $k \in \mathbb{Z}$ to its modulo- n residue class $[k]_n$. Now suppose that a polynomial $f \in \mathbb{Z}[x]$ is reducible over $\mathbb{Z}$; that is, there exist polynomials $g, h \in \mathbb{Z}[x]$ such that $f = gh$. Then $q_n(f) = q_n(g)q_n(h)$ is also reducible over $\mathbb{Z}_n$. Contrapositively, if $q_n(f)$ isn't reducible over $\mathbb{Z}_n$,

then f must be irreducible over $\mathbb{Z}$ (and hence also over $\mathbb{Q}$).

Example 11.3 Let $f = 2x^2 + 6x + 15$. By Eisenstein's criterion we can show this is irreducible over $\mathbb{Q}$ by considering $p = 3$.

Reducing modulo 7 gives $q_7(f) = 2x^2 + 6x + 1$, which is either irreducible, or factorises as two linear terms $(ax + b)(cx + d)$ with $a, b, c, d \in \mathbb{Z}_7$. Comparing coefficients, we see that $ac = 2$, $bd = 1$, and $(ad + bc) = 6$. A routine verification confirms that there are no modulo-7 solutions to these three equations, and therefore f is again irreducible over $\mathbb{Z}$ and thereby $\mathbb{Q}$.

For completeness, we state the following well-known definition:

Definition 11.4 Suppose that $f \in R[x]$ is a polynomial over some integral domain R . An element $\alpha \in R$ is a **root** or **zero** of f if $f(\alpha) = 0$.

By the Remainder Theorem,² the problem of finding roots of a polynomial is equivalent to that of factorising the polynomial into linear factors. For a quadratic polynomial

$$ax^2 + bx + c \quad (11.1)$$

the formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (11.2)$$

has been well known for centuries.³ The solutions given by this formula lie in $\mathbb{C}$, so the polynomial (11.1) can always be factorised into linear factors over $\mathbb{C}$, but not always over some subfield of $\mathbb{C}$.

Note, however, that Eisenstein's Criterion is a sufficient, but not a necessary condition for a polynomial $f \in \mathbb{Z}[x]$ to be irreducible. We'll now look at a couple of other useful tests for irreducibility, both of which are discussed in more detail in the article⁴ by Ram Murty.

Proposition 11.5 Let $f = a_mx^m + \cdots + a_1x + a_0 \in \mathbb{Z}[x]$, and set $N = \max\{|a_i/a_m| : 0 \leq i < m\}$. If $f(n)$ is prime for some integer $n > N+2$ then f is irreducible over $\mathbb{Z}$.

To prove this, we need the following lemma.

Lemma 11.6 Let $f = a_mx^m + \cdots + a_1x + a_0 \in \mathbb{Z}[x]$ and set $N = \max\{|a_i/a_m| : 0 \leq i < m\}$, then $|\alpha| < N+1$ for any root $\alpha \in \mathbb{C}$ of f .

Proof Since $f(\alpha) = a_m\alpha^m + \cdots + a_1\alpha + a_0 = 0$, we have

$$-a_m\alpha^m = a_{m-1}\alpha^{m-1} + \cdots + a_1\alpha + a_0$$

and so

$$|\alpha|^m \leq N(|\alpha|^{m-1} + \cdots + |\alpha| + 1) = N \left(\frac{|\alpha|^m - 1}{|\alpha| - 1} \right).$$

²Theorem 10.7, page 382.

³The earliest known examples of a general solution of what we would now regard as a quadratic equation occur in cuneiform tablets dated to the Sumerian Ur III period (c.2000BC), and the Berlin Papyrus, dated to the Middle Kingdom period (c.2050–c.1650BC) of Egyptian history. Further developments by Babylonian, Chinese, Indian and Greek mathematicians followed over the next three thousand years, and the Indian mathematician Brahmagupta (598–c.670AD) gave probably the first known explicit solution in a form we would recognise today. The first person to give a fully general formula to solve an arbitrary quadratic equation was probably the Flemish mathematician and engineer Simon Stevin (1548–1620) in 1594.

⁴M Ram Murty, *Prime numbers and irreducible polynomials*, The American Mathematical Monthly 109.5 (May 2002) 452–458.

If $|\alpha| \leq 1$ then $|\alpha| < N+1$ and the lemma follows; if $|\alpha| > 1$ then

$$(|\alpha| - 1)|\alpha|^m \leq N(|\alpha|^m - 1) < N|\alpha|^m,$$

so $|\alpha|^{m+1} < (N+1)|\alpha|^m$, and hence $|\alpha| < N+1$. $\square$

Proof of Proposition 11.5 Suppose that f is reducible in $\mathbb{Z}[x]$; that is, $f = gh$ for some polynomials

$$g = b_k x^k + \cdots + b_1 x + b_0 \quad \text{and} \quad h = c_l x^l + \cdots + c_1 x + c_0$$

in $\mathbb{Z}[x]$, both with positive degree. Since $f(n)$ is prime, either $g(n)$ or $h(n)$ must be equal to ± 1 . Suppose, without loss of generality, that $g(n) = \pm 1$. Then $g(x) = b_k \prod_i (x - \alpha_i)$, where i ranges over the indices of the complex zeros of g , which are themselves a subset of the zeros of f . By Lemma 11.6 we have

$$1 = |g(n)| \geq \prod_i (n - |\alpha_i|) > \prod_i (n - (N+1)) \geq 1,$$

which is a contradiction. Thus f must be irreducible. $\square$

Example 11.7 The polynomial $f = x^4 - 4x^2 + 16 \in \mathbb{Z}[x]$ cannot be shown to be irreducible by Eisenstein's Criterion: the only possible prime we can try is $p = 2$, and although conditions (i) and (ii) are satisfied, condition (iii) isn't.

Using Proposition 11.5, however, we just need to find a positive integer $n > (16+2) = 18$ such that $f(n)$ is prime. After a few attempts (which can be easily automated with a very short computer program) we find that $f(23) = 277741$, which is prime. Hence f is irreducible over $\mathbb{Z}$ (and therefore also over $\mathbb{Q}$).

A related irreducibility test is originally due to Arthur Cohn (1894–1940), a doctoral student of the algebraist Issai Schur (1875–1941) at the University of Berlin. Cohn's original version was stated in base 10,⁵ but it was later generalised to arbitrary bases.⁶

⁵ G Pólya and G Szegő, *Problems and Theorems in Analysis*, Springer (1976), volume II, page 133.

⁶ J Brillhart, M Filaseta, and A Odlyzko, *On an irreducibility theorem of A Cohn*, Canadian Journal of Mathematics 33 (1981) 1055–1059.

Proposition 11.8 (Cohn's Irreducibility Criterion) *If a prime integer p can be expressed in base- b notation for $b > 1$ as*

$$p = a_m b^m + a_{m-1} b^{m-1} + \cdots + a_1 b + a_0,$$

with $0 \leq a_i < b$ for $0 \leq i \leq m$, then the polynomial $f = a_m x^m + \cdots + a_1 x + a_0$ is irreducible over $\mathbb{Z}$.

The proof is somewhat involved, relying on a couple of technical lemmas, and would be a little too much of a digression here, but the details can be found in the article by Brillhart, Filaseta and Odlyzko. This article also contains a similar test that can be applied to polynomials with some negative coefficients, while Cohn's original test and the generalised version above can't.

Example 11.9 Let $f = x^4 + 6x^2 + 16$. Cohn's original, base-10 test is inapplicable, because the constant term $16 \geq 10$. But we can apply the generalised test and search for an integer $b > 16$ for which $f(b) = b^4 + 6b^2 + 16$ is prime. Again, this can be done very quickly with a short computer program, and we find that $b = 31$ works, since $f(31) = 929303$ is prime. Hence f is irreducible over $\mathbb{Z}$.

We can also use Cohn's Criterion to find irreducible polynomials:

Example 11.11 Expressing the prime $p = 97$ in different numerical bases $b > 1$, we obtain the following list of representations, which correspond to irreducible polynomials in $\mathbb{Z}[x]$.

$$\begin{array}{ll} 1100001_2 \longleftrightarrow x^6 + x^5 + 1, & 10121_3 \longleftrightarrow x^4 + x^2 + 2x + 1, \\ 1201_4 \longleftrightarrow x^3 + 2x^2 + 1, & 342_5 \longleftrightarrow 3x^2 + 4x + 2, \\ 241_6 \longleftrightarrow 2x^2 + 4x + 1, & 166_7 \longleftrightarrow x^2 + 6x + 6, \\ 141_8 \longleftrightarrow x^2 + 4x + 1, & 117_9 \longleftrightarrow x^2 + x + 7, \\ 97_{10} \longleftrightarrow 9x + 7, & 89_{11} \longleftrightarrow 8x + 9, \quad \dots \end{array}$$

As with Eisenstein's Criterion, there exist polynomials in $\mathbb{Z}[x]$ that are irreducible but which aren't detected by either Proposition 11.5 or Cohn's Criterion. For example, the polynomial $x^2 + 3x + 4$ is irreducible over $\mathbb{Z}$, but evaluates to an even (and thus non-prime) integer for any integral x .

Proposition 11.8 is a partial converse of Bunyakovsky's Conjecture, originally stated in 1857 by the Russian mathematician Viktor Bunyakovsky (1804–1889), and which is still an open question.

Conjecture 11.10 (Bunyakovsky's Conjecture) *Let $f \in \mathbb{Z}[x]$ be an irreducible, primitive polynomial of positive degree, with positive leading coefficient. Then there are infinitely many values of $n \in \mathbb{Z}$ for which $f(n)$ is prime.*

11.2 Field extensions

THE POLYNOMIAL $x^2 + 1$ is irreducible over $\mathbb{Z}$ (and hence also $\mathbb{Q}$). If, however, we switch to the larger field $\mathbb{Q}[i] = \{a + bi : a, b \in \mathbb{Q}\}$ then it factorises quite readily as $(x+i)(x-i)$. Similarly, the polynomial $x^2 - 2$ is irreducible over $\mathbb{Q}$, but not over $\mathbb{Q}[\sqrt{2}] = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$, in which it factorises as $(x + \sqrt{2})(x - \sqrt{2})$.

In some sense, $\mathbb{Q}[i]$ is the smallest possible field over which $x^2 + 1$ is reducible: no smaller subfield will do. Similarly, $\mathbb{Q}[\sqrt{2}]$ is the smallest field over which $x^2 - 2$ is reducible.

What we've done in both of these cases is to **adjoin** a carefully chosen element to $\mathbb{Q}$ in order to get a new field over which the polynomial is reducible. In general, however, we don't know that $\mathbb{Q}[\alpha]$ will be a field. It will certainly be an integral domain, but we also need it to have all necessary inverses.

The rational field $\mathbb{Q}$ has a privileged position in the theory we're developing in this chapter: it is the smallest nontrivial subfield of $\mathbb{C}$:

The proof of the transcendency of π will hardly diminish the number of circle-squarers, however; for this class of people has always shown an absolute distrust of mathematicians and a contempt for mathematics that cannot be overcome by any amount of demonstration.

— Felix Klein (1849–1925),
The Evanston Colloquium,

Lecture 7 (The transcendency of the numbers e and π), 4 September 1893



Wikimedia Commons / Anonymous Dutch artist

The Dutch mathematician and engineer **Simon Stevin** (c.1548–1620) is believed to have been born in Bruges in approximately 1548 to an unmarried couple named Anton Stevin and Catelyne van der Poort. Little is known of his early life, but after working as a clerk in Antwerp and Bruges, and travelling throughout Poland, Prussia and Norway during the early 1570s, he entered the University of Leiden in 1583 at the age of 35. During this period he became friends with Prince Maurits, Count of Nassau (1567–1625), whom he also tutored in mathematics. When Maurits' father William I, Prince of Orange, was assassinated in 1584, Maurits was appointed head of the army, and Stevin became one of his trusted advisors.

He wrote eleven books, including contributions to mathematics and engineering. In *De Thiende* ("The tenth"), published in 1585, he introduced decimal fractions to a Western audience for the first time. In his 1594 book *Arithmetica*, he presented the first known general solution for the quadratic equation.

In 1586 he devised an improved method of using windmills to pump water from marshes and flood plains that trebled the efficiency of the pumping process. In about 1600 he invented a land yacht: a sail-propelled chariot that could overtake galloping horses. Partly in honour of Stevin's achievement, mechanical engineering students at Eindhoven University of Technology regularly build, design and race land yachts at speeds of up to 70km/h.

Proposition 11.12 Let $F \subseteq \mathbb{C}$ be a nontrivial subfield of $\mathbb{C}$. Then $\mathbb{Q} \subseteq F$.

Proof Since F is a nontrivial field, it must contain 0 and 1 (which may not be equal) and hence we can inductively generate the positive integers $n = 1 + \cdots + 1$. Existence of additive inverses ensures that F also contains the negative integers, so $\mathbb{Z} \subseteq F$. The existence of multiplicative inverses ensures that F also contains rational numbers of the form $\frac{1}{n} = n^{-1}$ for all $n \in \mathbb{Z} \setminus \{0\}$, and together with the integers these multiplicatively generate the rest of the rational numbers, so $\mathbb{Q} \subseteq F$. $\square$

In the examples above, we constructed $\mathbb{Q}[i]$ and $\mathbb{Q}[\sqrt{2}]$ by **adjoining** an extra element i or $\sqrt{2}$. The point of this is that we've enlarged $\mathbb{Q}$ just enough to guarantee that the polynomial in question is reducible. More generally, we've defined a new field that contains $\mathbb{Q}$ as a subfield.

Definition 11.13 Let K and L be fields. Then L is an **extension** or **extension field** of K if there exists an injective field homomorphism $i: K \hookrightarrow L$ including K as a subfield of L . We denote this by $L:K$.

In particular, we're interested in finding out just how far we have to extend $\mathbb{Q}$ so that a given irreducible polynomial $f \in \mathbb{Q}[x]$ can be reduced. In the case of $x^2 + 1$ the corresponding extension is $\mathbb{Q}[i]:\mathbb{Q}$ and for $x^2 - 2$ it's $\mathbb{Q}[\sqrt{2}]:\mathbb{Q}$. More generally, we need the following definition.

Definition 11.14 Let F be a field. Then

$$F(x) = \{p/q : p, q \in F[x], q \neq 0\}$$

is the field of **rational functions** over F ; that is, the field of quotients $Q(F[x])$ of the polynomial ring $F[x]$.

Now let $\alpha \in K$ for some field K such that $F \subseteq K$. Then

$$F(\alpha) = \{p(\alpha)/q(\alpha) : p, q \in F[x], q(\alpha) \neq 0\}$$

is the field obtained by **adjoining** α to F .

It so happens that $\mathbb{Q}[i] = \mathbb{Q}(i)$ and $\mathbb{Q}[\sqrt{2}] = \mathbb{Q}(\sqrt{2})$, but it isn't necessarily the case that $F[\alpha] = F(\alpha)$. We will usually use the first of these to denote

$$F[\alpha] = \{a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_n\alpha^n : a_0, \dots, a_n \in F\}$$

which in general is only guaranteed to be a ring.

So, in order to obtain a field in which $x^2 + 1$ is reducible, we only have to adjoin i to $\mathbb{Q}$ to get $\mathbb{Q}(i)$. The field extension we're interested in is therefore $\mathbb{Q}(i):\mathbb{Q}$. Similarly, with $x^2 - 2$ we want to study the field extension $\mathbb{Q}(\sqrt{2}):\mathbb{Q}$.

Definition 11.15 Suppose that $L:K$ is a field extension with $L = K(\alpha)$ for some $\alpha \in L$. Then $L:K$ is called a **simple extension**, and α is said to be a **primitive element** of the extension.

We can also define $F(\alpha, \beta) = F(\alpha)(\beta)$; that is, the field obtained from F by first adjoining α and then adjoining β to the result. This enables us to define $F(S)$ for some finite set $S \subset K$.

Some simple field extensions might not look simple at first glance:

Example 11.16 The field extension $\mathbb{Q}(\sqrt{3}, i): \mathbb{Q}$ is simple, despite not initially appearing to be so. To see this, we have to find an element $\alpha \in \mathbb{Q}(\sqrt{3}, i)$ such that $\mathbb{Q}(\sqrt{3}, i) = \mathbb{Q}(\alpha)$. But $\alpha = (\sqrt{3} + i)$ is one such element. Certainly $\mathbb{Q}(\alpha) \subseteq \mathbb{Q}(\sqrt{3}, i)$, so we only need to show that $\mathbb{Q}(\sqrt{3}, i) \subseteq \mathbb{Q}(\alpha)$.

But it so happens that $i = \frac{1}{8}\alpha^3$ and hence $\sqrt{3} = \alpha - \frac{1}{8}\alpha^3$. Hence any element of $\mathbb{Q}(\sqrt{3}, i)$ can be obtained as a polynomial expression in α with rational coefficients, and so $\mathbb{Q}(\sqrt{3}, i) = \mathbb{Q}(\alpha)$.

Since $\frac{1}{8}\alpha^3 = i$, it is also the case that $\alpha^6 + 64 = 0$. Equivalently, α is a root of the polynomial $x^6 + 64 \in \mathbb{Q}[x]$. On the other hand, there is no polynomial $f \in \mathbb{Q}[x]$ such that $f(\pi) = 0$. This leads to a distinction between **algebraic** real numbers like $\sqrt{2}$ and $\sqrt[3]{5}$ (and, for that matter, 2 and $\frac{3}{7}$) that can be expressed as roots of polynomials in $\mathbb{Q}[x]$, and **transcendental** real numbers like π and e , that can't. We can generalise this distinction to arbitrary field extensions:

Definition 11.17 Let $L:K$ be a field extension, and suppose that $\alpha \in L$. Then α is **algebraic over K** if there exists a nonzero polynomial $f \in K[x]$ such that $f(\alpha) = 0$; otherwise we say that α is **transcendental over K** .

The extension $L:K$ is **algebraic** if every $\alpha \in L$ is algebraic over K , and **transcendental** otherwise.

The usual notion of algebraicity and transcendental fits into this definition: numbers such as π and e are transcendental over $\mathbb{Q}$, while numbers like $\sqrt{2}$, $\sqrt[3]{5}$ and 2 are algebraic over $\mathbb{Q}$.

Example 11.18 The extensions $\mathbb{Q}(i): \mathbb{Q}$, $\mathbb{Q}(\sqrt{2}): \mathbb{Q}$ and $\mathbb{Q}(\sqrt{3} + i): \mathbb{Q}$ are algebraic. However, the extension $\mathbb{R}: \mathbb{Q}$ is transcendental, because there exist some real numbers (for example, π and e) which aren't algebraic over $\mathbb{Q}$.

Given a field extension $L:K$, we can view L as a vector space over K in an obvious way: the elements of L are the 'vectors', and the elements of K are the 'scalars'. We can certainly add together any two elements of L , and L forms an abelian group under this 'vector addition' operation.



Wikimedia Commons

Charles Hermite (1822–1901) was the sixth of seven children of Ferdinand Hermite, a draper and artist. He studied at the prestigious École Polytechnique but was dismissed after a year, not for academic reasons but due to a disability of his right foot: the École was, and to some extent still is, a military academy. He completed his studies privately and pursued research independently, later returning to the École as a member of academic staff.

In 1858 he discovered a method for solving polynomial equations of degree 5 using elliptic functions, and in 1873 he proved that the number e is transcendental.



Wikimedia Commons

Carl Louis Ferdinand von Lindemann (1852–1939) was born in Hanover, the son of a schoolteacher. He studied mathematics at Göttingen, Erlangen and Munich, completing his doctorate on non-Euclidean geometry in 1873, under the supervision of Felix Klein (1849–1925).

His main contributions to mathematics were in geometry and analysis, and perhaps his best known achievement is his proof, published in 1882, that π is transcendental. This result was based on Charles Hermite's proof that e is transcendental, together with Euler's identity $e^{\pi i} = -1$.

In 1883 he was appointed to a chair at the University of Königsberg, and while there he supervised the doctoral dissertation of David Hilbert (1862–1943).

Also, since every element of K is contained in L , we can also multiply any element of L by an element of K , to get another element of L ; this 'scalar multiplication' operation also satisfies the usual associative and distributive laws, and K contains the multiplicative identity 1.

Definition 11.19 The **degree** $[L : K]$ of a field extension $L : K$ is the dimension $\dim_K L$ of L considered as a vector space over K .

For $L : K$ to have degree 1 requires L to be a 1-dimensional vector space over K , which is equivalent to saying $L \cong K$. Extensions of degree 1 are called **trivial extensions**. Extensions of degree 2 and 3 are called **quadratic** or **cubic** extensions, respectively. An extension $L : K$ is said to be **finite** or **infinite**, depending on whether its degree is finite or infinite.

For example, $\mathbb{Q}(i)$ has degree 2 over $\mathbb{Q}$. The reason for this is that, as noted earlier,

$$\mathbb{Q}(i) = \{a + bi : a, b \in \mathbb{Q}\}.$$

It is not difficult to see that $\mathbb{Q}(i)$ is a 2-dimensional rational vector space: the set $\{1, i\}$ is a suitable basis. Similarly, $\mathbb{Q}(\sqrt{2})$ also has degree 2 over $\mathbb{Q}$.

Example 11.20 The extension $\mathbb{Q}(\sqrt{3}, i) : \mathbb{Q}$ has degree 4. Recall that

$$\mathbb{Q}(\sqrt{3}, i) = \mathbb{Q}(\sqrt{3})(i) = \{p(i)/q(i) : p, q \in \mathbb{Q}(\sqrt{3})[x]\}.$$

By a rather tedious calculation it can be shown that

$$\mathbb{Q}(\sqrt{3}, i) = \{a + b\sqrt{3} + ci + di\sqrt{3} : a, b, c, d \in \mathbb{Q}\}.$$

The obvious basis for this is $\{1, \sqrt{3}, i, i\sqrt{3}\}$, which has four elements, and hence $[\mathbb{Q}(\sqrt{3}, i) : \mathbb{Q}] = \dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{3}, i) = 4$.

Also, note that $\mathbb{Q}(\sqrt{3}, i) = \{(a+b\sqrt{3}) + (c+d\sqrt{3})i : a, b, c, d \in \mathbb{Q}\}$, and so, considering $\mathbb{Q}(\sqrt{3}, i)$ as a vector space over $\mathbb{Q}(\sqrt{3})$ we see that $[\mathbb{Q}(\sqrt{3}, i) : \mathbb{Q}(\sqrt{3})] = \dim_{\mathbb{Q}(\sqrt{3})} \mathbb{Q}(\sqrt{3}, i) = 2$.

In the above example it turns out that

$$[\mathbb{Q}(\sqrt{3}, i) : \mathbb{Q}] = 4 = 2 \times 2 = [\mathbb{Q}(\sqrt{3}, i) : \mathbb{Q}(\sqrt{3})][\mathbb{Q}(\sqrt{3}) : \mathbb{Q}].$$

This is true in general: the degree is multiplicative, in much the same way that the index of a subgroup is.⁷ The formal statement of this fact is sometimes called the Tower Law.

Proposition 11.21 (The Tower Law) *Let $L : K$ and $K : F$ be finite field extensions. Then*

$$[L : F] = [L : K][K : F].$$

Proof Suppose that $[L : K] = m$ and $[K : F] = n$. Then K is an n -dimensional vector space over F , and so we can find a basis

$$A = \{a_1, \dots, a_n\}.$$

⁷ Proposition 2.32, page 55.

Similarly, L is an m -dimensional vector space over K , and so we can find a basis

$$B = \{b_1, \dots, b_m\}.$$

We claim that the mn products $a_i b_j$, where $1 \leq i \leq n$ and $1 \leq j \leq m$, form a basis for L over F .

First we need to show that they span L over F . Suppose that $c \in L$. Then there exist elements $\alpha_1, \dots, \alpha_n \in K$ such that

$$c = \sum_{i=1}^n \alpha_i a_i.$$

Furthermore, there exist elements $\beta_{i0}, \dots, \beta_{im} \in F$ such that

$$\alpha_i = \sum_{j=1}^m \beta_{ij} b_j$$

for $1 \leq i \leq n$. Hence

$$c = \sum_{i=1}^n \sum_{j=1}^m \beta_{ij} (a_i b_j)$$

and therefore $A \times B$ spans L over F .

To show linear independence, suppose that

$$\sum_{i=1}^n \sum_{j=1}^m \gamma_{ij} (a_i b_j) = 0$$

where $\gamma_{ij} \in F$. Then

$$\sum_{i=1}^n \left(\sum_{j=1}^m \gamma_{ij} b_j \right) a_i = 0$$

and since $a_1, \dots, a_n$ are linearly independent, it follows that

$$\sum_{j=1}^m \gamma_{ij} b_j = 0$$

for $1 \leq i \leq n$. And since $b_1, \dots, b_m$ are linearly independent, it follows that each $\gamma_{ij} = 0$, so all the products $a_i b_j$ are linearly independent, and hence they form a basis for L over F . Since this basis has mn elements, it follows that $[L : F] = \dim_F L = mn = [L : K][K : F]$. $\square$

The following corollary can be proved by a simple inductive argument.

Corollary 11.22 Suppose that $K_0 \subseteq K_1 \subseteq \dots \subseteq K_n$ are fields, such that $K_{i+1} : K_i$ are extensions for $0 \leq i < n$. Then

$$[K_n : K_0] = [K_n : K_{n-1}][K_{n-1} : K_{n-2}] \dots [K_1 : K_0].$$

All of the finite extensions we've seen so far have been algebraic. The next proposition shows that all finite extensions are algebraic:

Proposition 11.23 Any finite extension $L : K$ is algebraic.

Proof Let $\alpha \in L$ be some arbitrary element of the larger field L , and suppose that $[L:K] = n$. Recall that any m elements in an n -dimensional vector space cannot be linearly independent if $m > n$. In particular, the $n+1$ elements

$$1, \alpha, \alpha^2, \dots, \alpha^n$$

can't be linearly independent, and hence there exist $a_0, \dots, a_n \in K$, not all zero, such that

$$a_0 + a_1\alpha + a_2\alpha^2 + \dots + a_n\alpha^n = 0.$$

Hence $f = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ is a nonzero polynomial in $K[x]$ such that $f(\alpha) = 0$. Therefore α is algebraic over K . But α was arbitrarily chosen, so every element of L is algebraic over K , and thus $L:K$ is algebraic. $\square$

Corollary 11.24 *If $L:K$ is a transcendental extension, then it has infinite degree.*

In Example 11.16 we saw that some field extensions are simple even though they might not look like it at first glance. The next example describes a finite field extension that isn't simple.

Example 11.25 Let p be prime, and suppose that $\mathbb{F}_p$ is the finite field of order p . Then $L = \mathbb{F}_p(x, y)$ is the field of rational functions in two variables, with coefficients in $\mathbb{F}_p$. Now let $K = \mathbb{F}_p(x^p, y^p)$, the field of rational functions in x^p and y^p . Then K is a subfield of L , so $L:K$ is an extension of K .

The extension $\mathbb{F}_p(x^p, y) : \mathbb{F}_p(x^p, y^p)$ has degree p , and so does the extension $\mathbb{F}_p(x, y^p) : \mathbb{F}_p(x^p, y^p)$. So by Proposition 11.21 the extension $L:K$ has degree p^2 .

Now suppose that $f \in L = \mathbb{F}_p(x, y)$. Then $f = g/h$, where $g, h \in \mathbb{F}_p[x, y]$ and $h \neq 0$. By the Freshman's Binomial Theorem (Proposition 8.20) both g^p and h^p lie in $\mathbb{F}_p[x^p, y^p]$, and so $f = g/h$ lies in $\mathbb{F}_p(x^p, y^p) = K$. So every element of L satisfies a degree- p polynomial in K , and hence there is no single element of degree p^2 that generates L over K . Therefore $L:K$ is not simple.

The extension obtained from the field of rational functions is simple but transcendental.

Proposition 11.26 *For any field F , the extension $F(x):F$ is simple and transcendental.*

Proof The extension $F(x):F$ is simple by definition: the field of rational functions over F is obtained by adjoining a single element x to F . For this extension to be algebraic, we require that for every rational function $f \in F(x)$ there exists a nonzero polynomial $g \in F[x]$

such that $g(f) = 0$. But this certainly isn't the case for $f = x$, since any polynomial satisfying $g(x) = 0$ must be zero, and hence x is not algebraic over F in $F(x)$. $\square$

It turns out that all simple transcendental field extensions are of this form, although to properly state and prove that fact we first need to formulate a concept of equivalence of field extensions. Recall that an extension of a field F is an embedding of F into some larger field E . Given two such inclusions $i_1: F \hookrightarrow E_1$ and $i_2: F \hookrightarrow E_2$, what we're looking for is an isomorphism $\phi: E_1 \rightarrow E_2$ that respects the embedded images of F .

Definition 11.27 Let $E_1: F_1$ and $E_2: F_2$ be two field extensions determined by inclusion homomorphisms $i_1: F_1 \hookrightarrow E_1$ and $i_2: F_2 \hookrightarrow E_2$. Then a **homomorphism** of field extensions is a field homomorphism $\phi: E_1 \rightarrow E_2$ such that $\phi(i_1(F_1)) \subseteq i_2(F_2)$; that is, ϕ maps the embedded image of F_1 in E_1 to the embedded image of F_2 in E_2 . If ϕ is bijective and the restriction $\phi|_{i_1(F_1)}$ maps the image $i_1(F_1)$ isomorphically to the image $i_2(F_2)$ then we call it an **isomorphism** of field extensions, and say the extensions $E_1: F_1$ and $E_2: F_2$ are **isomorphic**.

Let $E_1: F$ and $E_2: F$ be two extensions of the same field F , determined by inclusion homomorphisms $i_1: F \hookrightarrow E_1$ and $i_2: F \hookrightarrow E_2$. An **F -homomorphism** is a field homomorphism $\phi: E_1 \rightarrow E_2$ such that $\phi(i_1(a)) = i_2(a)$ for all $a \in F$; that is, ϕ restricts to the identity map id_F on the embedded images of F . An **F -isomorphism** is a bijective F -homomorphism, and two extensions $E_1: F$ and $E_2: F$ are **F -isomorphic** if there is an F -isomorphism $\phi: E_1 \rightarrow E_2$.

In most of what follows, particularly in the classification of simple field extensions, and also when we discuss field automorphisms and the Galois group in Section 11.4, we will be mainly concerned with determining when two extensions of the same field F are F -isomorphic, but we will require the more general notion when we classify finite fields in Section 11.3.

The next proposition, together with Proposition 11.26, provides a neat classification of simple transcendental field extensions:

Proposition 11.28 Any simple transcendental extension $F(\alpha): F$ is isomorphic to $F(x): F$.

Proof Let $\phi: F(x) \rightarrow F(\alpha)$ be the evaluation homomorphism ev_α ,⁸ that is, $\phi(f/g) = f(\alpha)/g(\alpha)$ for any rational function $f/g \in F(x)$. Since g is a nonzero polynomial in $F[x]$ and α is transcendental over F , it follows that $g(\alpha) \neq 0$, so ϕ is a well-defined homomorphism. Furthermore, if $f(\alpha)/g(\alpha) = 0$ then we must have $f(\alpha) = 0$, which can only happen if $f(\alpha) = 0$, again because α is transcendental over

⁸ Example 9.20, page 353.

F . Therefore $\ker \phi = \{0\}$ and so ϕ is injective; it is clearly surjective as well, and is hence an isomorphism $\phi: F(x) \rightarrow F(\alpha)$. The restriction $\phi|_F$ is the identity, so ϕ is an isomorphism of extensions. $\square$

To classify simple algebraic extensions, we need to introduce the concept of a **minimal polynomial**. Returning again to Example 11.16, we found a polynomial $x^6 + 64$ with the required root $\alpha = \sqrt{3} + i$. But we can factorise this as

$$x^6 + 64 = (x^2 + 4)(x^4 - 4x^2 + 16)$$

where the first factor has roots $\pm 2i$ and the second factor has roots $\pm\sqrt{3} \pm i$. So if we're only interested in recovering α , we don't need to bother with the first factor $x^2 + 4$. In fact, $f = x^4 - 4x^2 + 16$ happens to be the lowest-degree polynomial in $\mathbb{Q}[x]$ with α as a root. We saw in Example 11.7 that this polynomial is irreducible over $\mathbb{Z}$ and $\mathbb{Q}$.

Proposition 11.29 *If α is algebraic over a field F , then there is a unique monic, irreducible polynomial (the **minimal polynomial**) of minimal degree in $F[x]$ such that $f(\alpha) = 0$, and $f|g$ for any other $g \in F[x]$ with $g(\alpha) = 0$.*

(Recall that a polynomial is **monic** if its highest-degree coefficient is 1.)

Proof If α is algebraic over F then by definition there exists a polynomial $f \in F[x]$ such that $f(\alpha) = 0$. We may suppose that f is monic; if not, then we can obtain a new polynomial by dividing each coefficient by the highest-degree coefficient, and this new polynomial will also have α as a root. Furthermore, suppose that f has minimal degree over all monic polynomials in $F[x]$ with α as a root.

This monic polynomial f must be unique: if not, then there is another monic polynomial $h \in F[x]$ with $h(\alpha) = 0$. Then $f(\alpha) - h(\alpha) = 0$, which means that either $f = h$ or we can find a monic polynomial $k \in F[x]$ that is a constant multiple of $f - h$, and $k(\alpha) = 0$. $\square$

The minimal polynomial of an algebraic extension encodes important information about the extension. In particular, the degree of the extension is equal to the degree of its minimal polynomial:

Proposition 11.30 *If $F(\alpha) : F$ is a simple algebraic extension of F , then $[F(\alpha) : F] = \deg(f)$, where f is the minimal polynomial of α over F .*

Proof If α is algebraic over F then by Proposition 11.29 there exists a nonzero, monic minimal polynomial

$$f = a_0 + a_1x + \cdots + x^n \in F[x]$$

such that $f(\alpha) = 0$. Here $\deg(f) = n$. We claim that any element $b \in F(\alpha)$ can be expressed uniquely as an F -linear combination

$$b = b_0 + b_1\alpha + b_2\alpha^2 + \cdots + b_{n-1}\alpha^{n-1}$$

where $b_0, \dots, b_{n-1} \in F$. That is, the set $\{1, \alpha, \alpha^2, \dots, \alpha^{n-1}\}$ is linearly independent in $F(\alpha)$, and hence forms a basis for $F(\alpha)$ over F .

Since $f(\alpha) = 0$, it follows that $a_0 + a_1\alpha + \dots + \alpha^n = 0$, and hence

$$\alpha^n = -a_0 - a_1\alpha - \dots - a_{n-1}\alpha^{n-1}.$$

Therefore, any monomial of the form α^k can be expressed as an F -linear combination of powers of α with exponent strictly less than n , and hence $1, \dots, \alpha^{n-1}$ span $F(\alpha)$ over F .

To show linear independence, suppose that

$$b_0 + b_1\alpha + b_2\alpha^2 + \dots + b_{n-1}\alpha^{n-1} = 0$$

with at least one coefficient not equal to zero. Then

$$\frac{b_0}{b_{n-1}} + \frac{b_1}{b_{n-1}}\alpha + \dots + \frac{b_{n-2}}{b_{n-1}}\alpha^{n-2} + \alpha^{n-1} = 0.$$

Suppose that b_k is the highest nonzero coefficient. Then

$$g(x) = \frac{b_0}{b_k} + \frac{b_1}{b_k}x + \dots + \frac{b_{k-1}}{b_k}x^{k-1} + x^k$$

is a monic polynomial over F with $g(\alpha) = 0$, with degree $k < n$. But this contradicts the assertion that the minimal polynomial f , with degree n , was the smallest-degree monic, irreducible polynomial over F with α as a root. Hence $b_0 = \dots = b_{n-1} = 0$, and therefore $1, \alpha, \dots, \alpha^{n-1}$ are linearly independent. Thus they form a basis for $F(\alpha)$ over F and hence $[F(\alpha) : F] = n$. $\square$

By Proposition 11.23 we know that finite extensions are algebraic, and therefore (by Corollary 11.24) that transcendental extensions must be infinite. However, not every algebraic extension is finite:

Example 11.31 The extension $\mathbb{Q}(2^{1/2}, 2^{1/4}, 2^{1/8}, \dots) : \mathbb{Q}$ is algebraic but infinite. It is algebraic, because each $2^{1/2^n}$ is algebraic over $\mathbb{Q}$, and therefore over $\mathbb{Q}(2^{1/2}, 2^{1/4}, 2^{1/8}, \dots, 2^{1/2^{n-1}})$, since for each $n \in \mathbb{N}$ we can set $f_n = x^{2^n} - 2 \in \mathbb{Q}[x]$ and then $f_n(2^{1/2^n}) = 0$.

This extension is infinite, because the infinite set $\{2^{1/2^n} : n \in \mathbb{N}\}$ is linearly independent over $\mathbb{Q}$.

Other examples of infinite algebraic extensions include

$$\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5}, \dots) : \mathbb{Q},$$

obtained by adjoining all square roots of positive primes,

$$\mathbb{Q}(\omega_3, \omega_4, \omega_5, \dots) : \mathbb{Q},$$

where ω_n is a primitive n th root of unity (that is, a minimal $\omega_n \neq 1$ such that $\omega_n^n = 1$) and $\mathbb{A} : \mathbb{Q}$, where

$$\mathbb{A} = \{x \in \mathbb{C} : x \text{ is algebraic over } \mathbb{Q}\}$$

is the field of **algebraic numbers**.



Wikimedia Commons

Born in Prussia to a wealthy Jewish family, **Leopold Kronecker** (1823–1891) studied mathematics, astronomy and philosophy at the University of Berlin, graduating in 1845 with a dissertation on algebraic number theory prepared under the supervision of Peter Gustav Lejeune Dirichlet (1805–1859).

He spent the next ten years managing his family's large farming estate, studying and researching mathematics in his spare time, and returned to Berlin in 1855, publishing numerous papers based on his earlier research.

He was elected to the Prussian Academy of Sciences in 1861 and taught at Berlin from 1862 onwards, where he remained for the rest of his career, turning down the offer of a chair at Göttingen in 1866.

He made many contributions to number theory and algebra. In 1850 he devised a general solution to the quintic equation using group-theoretic methods (although not in terms of radicals, for reasons we'll see in a little while).

He was a leading member of the **constructivist** movement, which held that a mathematical object did not exist unless it could actually be constructed; in particular, proofs by contradiction weren't considered valid.

His mathematical disputes tended to drift into personal animosity, and some of his contemporaries found him difficult to work with, notably Karl Weierstrass (1815–1897) who nearly resigned his chair in 1885. He also had a bitter and long-running conflict with Georg Cantor (1845–1918) over the latter's work on transfinite numbers.

Going back to our key examples $x^2 - 2$ and $x^2 + 1$, the point of all of this theory on field extensions is that we can factorise $x^2 - 2 = (x + \sqrt{2})(x - \sqrt{2})$ over $\mathbb{Q}(\sqrt{2})$ and $x^2 + 1 = (x + i)(x - i)$ over $\mathbb{Q}(i)$. In both cases, we adjoin a carefully chosen element to $\mathbb{Q}$ to obtain a field over which the given polynomial is reducible; that is, it factorises as a product of non-constant polynomials of lower degree. In these two cases, the polynomials factorise as linear terms, but this is not always the case for polynomials of cubic or higher degree.

For example, let $f = x^3 + 5$. This is irreducible over $\mathbb{Q}$ (to see this, apply Eisenstein's Criterion with $p = 5$). If we adjoin $\alpha = \sqrt[3]{5}$ then we find that f is reducible over $\mathbb{Q}(\alpha)$:

$$f = x^3 - 5 = (x - \alpha)(x^2 + \alpha x + \alpha^2).$$

But $x^2 + \alpha x + \alpha^2$ doesn't factorise as linear factors over $\mathbb{Q}(\alpha)$. For that, we have to adjoin another element: $\beta = i\sqrt{3}$ will do. Then f factorises over $\mathbb{Q}(\alpha, \beta)$ as

$$f = (x - \alpha)(x - \frac{\alpha}{2}(1 + \beta))(x - \frac{\alpha}{2}(1 - \beta)).$$

The degree of this extension $\mathbb{Q}(\alpha, \beta) : \mathbb{Q}$ is 6, since $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$ and $[\mathbb{Q}(\alpha, \beta) : \mathbb{Q}(\alpha)] = 2$.

Before we proceed any further with this line of inquiry, there's something we've taken for granted so far, and which we must now deal with. We've assumed that for a given polynomial f over some field F it's actually possible to find a suitably adjoinable element. By Definition 11.17 α is algebraic over F if there exists a finite polynomial over F with α as a root, and Proposition 11.29 says that we can always find a monic, irreducible polynomial (the **minimal polynomial**) satisfying this condition, and whose degree, by Proposition 11.30, is equal to the degree of the simple extension $[F(\alpha) : F]$.

What we haven't yet proved is that this works the other way round: given some irreducible polynomial $f \in F[x]$ we can always find some appropriate element α to adjoin to F , so that f is reducible over $F(\alpha)$. The confirmation of this is due to the German mathematician Leopold Kronecker (1823–1891).

Theorem 11.32 (Kronecker's Theorem) *Let F be a field, and let $f \in F[x]$ be an irreducible polynomial of degree $n > 0$. There exists a simple field extension $F(\alpha) : F$ of degree $[F(\alpha) : F] = n$ such that $f(\alpha) = 0$.*

Proof Let $i : F \hookrightarrow F[x]$ be the obvious inclusion homomorphism, mapping each element of F to the corresponding constant polynomial in $F[x]$. Now let $K = F[x]/\langle f \rangle$. Since f is irreducible in $F[x]$, then $\langle f \rangle$ is a maximal principal ideal by Proposition 10.25 (i). But $F[x]$ is a PID by Proposition 10.47, so $\langle f \rangle$ is a maximal ideal. Hence the quotient $F[x]/\langle f \rangle$ is a field by Proposition 9.61.

If $q: F[x] \rightarrow K$ is the quotient homomorphism, then the composite $q \circ i: F \rightarrow K$ is an injective field homomorphism. This is because any element $a \in F$ maps to $a + \langle f \rangle$ in K , and hence for any $a, b \in F$ with $(q \circ i)(a) = a + \langle f \rangle = b + \langle f \rangle = (q \circ i)(b)$ it follows that $a = b$ since $a, b \notin \langle f \rangle$. Therefore $K: F$ is a field extension.

Now let $\alpha = q(x) = x + \langle f \rangle$. Since x generates $F[x]$ over F , it follows that α generates K over F , and so $K = F(\alpha)$, so $K: F = F(\alpha): F$ is a simple field extension.

Suppose $f(x) = a_n x^n + \cdots + a_1 x + a_0$ for some $a_0, \dots, a_n \in F$. Then

$$\begin{aligned} f(\alpha) &= f(x + \langle f \rangle) \\ &= a_n(x + \langle f \rangle)^n + \cdots + a_1(x + \langle f \rangle) + a_0 \\ &= a_n(x^n + \langle f \rangle) + \cdots + a_1(x + \langle f \rangle) + a_0 \\ &= (a_n x^n + \cdots + a_1 x + a_0) + \langle f \rangle \\ &= f + \langle f \rangle \\ &= 0 + \langle f \rangle \end{aligned}$$

Hence $f(\alpha) = 0$ in K , and so α is algebraic over F . By Proposition 11.29 f must be a constant multiple of the minimal polynomial of α over F . Hence, by Proposition 11.30 $[K: F] = [F(\alpha): F] = \deg(f) = n$. $\square$

We can extend this to account for reducible polynomials too:

Corollary 11.33 *Let $f \in F[x]$ be a nonconstant polynomial of degree $n > 0$. Then there exists a simple field extension $F(\alpha): F$ of degree $[F(\alpha): F] \leq n$ such that $f(\alpha) = 0$.*

Proof If f is irreducible then the result follows immediately from the main theorem. Suppose instead that f is reducible. By Proposition 10.47, $F[x]$ is a PID and hence a UFD by Corollary 10.43. So f has a unique factorisation as a product of irreducible polynomials, each with degree strictly less than n . Let g be one of these factors, and suppose that $\deg(g) = m < n$. Then by the main theorem there exists a simple field extension $F(\alpha): F$ with $[F(\alpha): F] = m < n$ and $f(\alpha) = g(\alpha) = 0$. $\square$

Classifying simple algebraic extensions requires a little more work than transcendental ones, but the end result is still quite neat. We start with the following proposition, which characterises simple algebraic extensions in terms of quotients of polynomial rings.

Proposition 11.34 *Let $F(\alpha): F$ be a simple algebraic extension of a field F , and suppose that f is the minimal polynomial of α over F . Then $F(\alpha) \cong F[x]/\langle f \rangle$. The isomorphism can be chosen so as to map α to $x + \langle f \rangle$, and to be the identity when restricted to $F \subseteq F(\alpha)$.*

Proof Let $\theta: F[x] \rightarrow F(\alpha)$ by $g \mapsto g(\alpha)$; this is a surjective homomorphism. Its kernel $\ker \theta$ consists of all polynomials in $F[x]$ for which α is a zero. By Proposition 11.29 any such polynomial must have the minimal polynomial f as a factor. Hence $\langle f \rangle \subseteq \ker \theta$. But by Proposition 10.25 (i), $\langle f \rangle$ is maximal in $F[x]$ because f is irreducible. And since $\ker \theta$ doesn't contain the constant polynomial $1 \in F[x]$ we can't have $\ker \theta = F[x]$, so $\ker \theta = \langle f \rangle$.

⁹Theorem 9.49, page 364.

By the First Isomorphism Theorem for Rings,⁹ then, we have

$$F[x]/\langle f \rangle = F[x]/\ker \theta \cong \text{im } \theta = F(\alpha)$$

as claimed. This isomorphism is the induced map $\hat{\theta}: F[x]/\langle f \rangle \rightarrow F(\alpha)$, which maps $x + \langle f \rangle$ to α and any constant element $a + \langle f \rangle$ to $a \in F(\alpha)$. Its inverse $\hat{\theta}^{-1}: F(\alpha) \rightarrow F[x]/\langle f \rangle$ is the required isomorphism. $\square$

The next proposition is slightly more general than we need to classify simple algebraic field extensions: all we need at this point is the subsequent corollary, but we'll need the full version later so we might as well do it here.

There is a slight abuse of notation here, in that we're using ϕ to denote both the field isomorphism $\phi: F_1 \rightarrow F_2$, and also the corresponding ring isomorphism $\phi: F_1[x] \rightarrow F_2[x]$. This shouldn't present a problem in practice, but it's worth being aware of it anyway.

Proposition 11.35 *Let $\phi: F_1 \rightarrow F_2$ be a field isomorphism, and suppose that $f \in F_1[x]$ is irreducible over F_1 . Let α be a root of f in some extension of F_1 , and let β be a root of $\phi(f)$ in some extension of F_2 . Then there is an isomorphism $\psi: F_1(\alpha) \rightarrow F_2(\beta)$ that maps α to β and agrees with ϕ on F_1 ; that is, $\psi|_{F_1} = \phi$.*

Proof By Proposition 11.34, there exist isomorphisms

$$\theta_\alpha: F_1(\alpha) \rightarrow F_1[x]/\langle f \rangle \quad \text{and} \quad \theta_\beta: F_2(\beta) \rightarrow F_2[x]/\langle \phi(f) \rangle,$$

such that $\theta_\alpha(\alpha) = x + \langle f \rangle$ and $\theta_\beta(\beta) = x + \langle \phi(f) \rangle$.

The composite $\psi = (\theta_\beta^{-1} \circ \phi \circ \theta_\alpha): F_1(\alpha) \rightarrow F_2(\beta)$ thus maps α to β , and restricts to ϕ on F_1 . $\square$

Corollary 11.36 *Let $F(\alpha): F$ and $F(\beta): F$ be simple algebraic extensions of a field F where α and β have the same minimal polynomial f over F . Then $F(\alpha) \cong F(\beta)$, and the isomorphism can be chosen to map α to β . Furthermore, the simple algebraic extensions $F(\alpha): F$ and $F(\beta): F$ are F -isomorphic.*

Proof This follows immediately from Proposition 11.35 by setting $F_1 = F_2 = F$ and $\phi = \text{id}_F$. The resulting isomorphism $\psi: F(\alpha) \rightarrow F(\beta)$ restricts to the identity on F , and is hence an F -isomorphism. $\square$

This completes our classification. Every nonconstant polynomial $f \in F[x]$ yields a simple algebraic field extension $F(\alpha): F$ where α is a root of f . And any two simple algebraic extensions $F(\alpha): F$ and $F(\beta): F$ are isomorphic if α and β have the same minimal polynomial over F .

For example, both $-\sqrt{2}$ and $\sqrt{2}$ have minimal polynomial $x^2 - 2$ over $\mathbb{Q}$, so by Proposition 11.35 the extensions $\mathbb{Q}(\sqrt{2}) : \mathbb{Q}$ and $\mathbb{Q}(-\sqrt{2}) : \mathbb{Q}$ are $\mathbb{Q}$ -isomorphic via an isomorphism $\phi : \mathbb{Q}(\sqrt{2}) \rightarrow \mathbb{Q}(-\sqrt{2})$ that switches $\sqrt{2}$ and $-\sqrt{2}$, and restricts to the identity on $\mathbb{Q}$.

By the Remainder Theorem,¹⁰ the existence of a root α of a polynomial f is equivalent to f having a linear factor $(x - \alpha)$. So, now we'd like to study factorisations of polynomials into linear factors, and thus we introduce some more new terminology.

¹⁰ Theorem 10.7, page 382.

Definition 11.37 If a polynomial $f \in F[x]$ can be factorised as a product of linear factors

$$f = (x - \alpha_1)(x - \alpha_2) \dots (x - \alpha_n)$$

where $\alpha_1, \dots, \alpha_n \in K$, where K is some field such that $F \subseteq K$, then we say that f **splits** over K .

Splittability is a stronger condition than reducibility: if a polynomial is reducible then it factorises into lower-degree terms, but if it's splittable then it factorises completely as a product of linear terms. Given some polynomial $f \in F[x]$, we want to find the simplest possible extension of F in which f is not only reducible, but splits:

Definition 11.38 Suppose that $f \in F[x]$ splits over some field K for which $F \subseteq K$, and that f does not split over any proper subfield of K . Then K is a **splitting field** for f over F , and $K : F$ is a **splitting field extension** for f .

We'll start by looking at two of our key examples.

Example 11.39 The splitting field for $f = x^2 - 2$ over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt{2})$, in which f factorises as $(x + \sqrt{2})(x - \sqrt{2})$. There is no smaller field in which f splits. Furthermore, f is the minimal polynomial of this extension, so $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = \deg(f) = 2$.

Example 11.40 The splitting field for $f = x^2 + 1$ over $\mathbb{R}$ is the complex number field $\mathbb{C} = \mathbb{R}(i)$, which has degree 2.

The next example is slightly more complicated, and is one that we'll return to later when we discuss the Galois Correspondence.

Example 11.41 Let's investigate the splitting field for $f = x^3 - 5$ over $\mathbb{Q}$. Set $\alpha = \sqrt[3]{5}$ and $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. Then the roots of f are α , $\alpha\omega$ and $\alpha\omega^2$. The splitting field of f over $\mathbb{Q}$ is the smallest field extension of $\mathbb{Q}$ in which f splits. It must therefore contain α and ω , which are linearly independent over $\mathbb{Q}$, so the smallest such extension field is $\mathbb{Q}(\alpha, \omega)$.

We can calculate the degree of this extension using the Tower Law:

$[\mathbb{Q}(\alpha, \omega) : \mathbb{Q}] = [\mathbb{Q}(\alpha, \omega) : \mathbb{Q}(\alpha)][\mathbb{Q}(\alpha) : \mathbb{Q}]$. The minimal polynomial of α over $\mathbb{Q}$ is $x^3 - 5$, which has degree 3, so $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$. The minimal polynomial of ω over $\mathbb{Q}$ (and also over $\mathbb{Q}(\alpha)$) is $x^2 + x + 1$, which has degree 2, so $[\mathbb{Q}(\alpha, \omega) : \mathbb{Q}(\alpha)] = 2$, and by the Tower Law $[\mathbb{Q}(\alpha, \omega) : \mathbb{Q}] = 2 \times 3 = 6$.

The degree of $\mathbb{Q}(\alpha, \omega) : \mathbb{Q}$ is the dimension of $\mathbb{Q}(\alpha, \omega)$ as a vector space over $\mathbb{Q}$. A suitable basis is $\{1, \alpha, \alpha^2, \omega, \alpha\omega, \alpha\omega^2\}$.

Splitting fields exist and have finite degree:

Proposition 11.42 *Given a field F and a polynomial $f \in F[x]$, there exists a splitting field K for f . If $\deg(f) = n$ then $[K : F] \leq n!$*

Proof We prove this by induction on $n = \deg(f)$. If $n = 1$ then f is linear and therefore splits over F , so $K = F$ and we have $[K : F] = 1$.

Now suppose that $n \geq 2$, and that the hypothesis holds for all polynomials in $F[x]$ with degree strictly less than n .

If f is reducible over F then we can factorise it as $f = gh$ with $\deg(g) = p < n$ and $\deg(h) = q < n$. By induction there is a splitting field L for g over F , with $[L : F] \leq p!$ and over L g factorises as

$$g = a(x - \alpha_1) \dots (x - \alpha_p)$$

where $a \in F$ and $\alpha_1, \dots, \alpha_p \in L$. Furthermore, $L = F(\alpha_1, \dots, \alpha_p)$.

Now $h \in F[x] \subseteq L[x]$ so we can regard h as a polynomial over L . Again, by induction there exists a splitting field K for h over L , with $[K : L] \leq q!$ and we can factorise h as

$$h = b(x - \beta_1) \dots (x - \beta_q)$$

for some $b \in L$ and $\beta_1, \dots, \beta_q \in K$. Moreover, $K = L(\beta_1, \dots, \beta_q) = F(\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q)$. Thus K is a splitting field for f over F and by the Tower Law¹¹ we have

$$[K : F] = [K : L][L : F] \leq q!p! \leq (p + q)! = n!$$

as claimed.

If f is irreducible over F then by Kronecker's Theorem we can find a simple extension $F(\alpha) : F$ with degree $[F(\alpha) : F] = \deg(f) = n$ such that $f(\alpha) = 0$. Therefore $f = (x - \alpha)g$ for some polynomial $g \in F[x]$ with $\deg(g) = n - 1$.

By the inductive hypothesis there exists a splitting field K for g over $F(\alpha)$ with degree $[K : F(\alpha)] \leq (n - 1)!$ and g factorises over K as

$$g = a(x - \beta_1) \dots (x - \beta_{n-1})$$

for some $\beta_1, \dots, \beta_{n-1} \in K = F(\alpha)(\beta_1, \dots, \beta_{n-1}) = F(\alpha, \beta_1, \dots, \beta_{n-1})$ and $a \in F$. Thus $K : F$ is a splitting field extension for f . $\square$

Furthermore, splitting fields are unique up to isomorphism.

¹¹ Proposition 11.21, page 428.

Proposition 11.43 *Let $\phi: F_1 \rightarrow F_2$ be an isomorphism of fields, and suppose that $f \in F_1[x]$. Suppose also that E_1 is a splitting field for f over F_1 , and that E_2 is a splitting field for $\phi(f)$ over F_2 . Then there is an isomorphism $\psi: E_1 \rightarrow E_2$ which agrees with ϕ on F_1 ; that is, $\psi|_{F_1} = \phi$.*

Proof We prove this by induction on $\deg(f)$. If $\deg(f) = 1$ then f is linear and already splits over F_1 , so $E_1 = F_1$, $E_2 = F_2$ and $\psi = \phi$.

Now suppose that $\deg(f) = n > 1$, and that the proposition has been proved for all polynomials of degree less than n .

Let $g \in F_1[x]$ be a monic, irreducible factor of f . Let α be a root of g in E_1 , and let β be a root of $\phi(g)$ in E_2 . Then g is the minimal polynomial of the simple algebraic extension $F(\alpha): F$

By Proposition 11.35, there is an isomorphism $\psi: F_1(\alpha) \rightarrow F_2(\beta)$ that agrees with ϕ when restricted to F_1 , and which maps α to β .

We can now factorise $f = (x - \alpha)h$ for some $h \in F_1(\alpha)[x]$. Then E_1 is a splitting field for h over $F_1(\alpha)$ and E_2 is a splitting field for $\psi(h)$ over $F_2(\beta)$. By the inductive hypothesis, since $\deg(h) < \deg(f)$, it follows that there is an isomorphism $\chi: E_1 \rightarrow E_2$ that agrees with ψ on $F_1(\alpha)$ and hence with ϕ when restricted to F_1 . $\square$

Corollary 11.44 *Let f be a nonconstant polynomial in $F[x]$ and suppose that E_1 and E_2 are two splitting fields for f over F . Then there is an F -isomorphism $\psi: E_1 \rightarrow E_2$.*

Proof This follows immediately from the above proposition by setting $F_1 = F_2 = F$ and $\phi = \text{id}_F$. $\square$

11.3 Finite fields

WE SAW EARLIER that, for any prime p , the set $\mathbb{F}_p = \{0, \dots, p-1\}$ forms a field under addition and multiplication modulo p . At the time we also claimed that there is a single field of order $q = p^n$ for any $n \in \mathbb{N}$. We can now prove this. First let's look at an illustrative example.

Example 11.45 One of the rings we saw in Example 8.7 was a field of order 4: the ring $\mathbb{Z}_2[\omega]$, where $\omega = \frac{1}{2} + \frac{\sqrt{3}}{2}i$ is a primitive cube root of unity; that is, a nontrivial root of the polynomial $x^3 - 1$. The addition and multiplication tables for this ring are shown in Table 11.1. Examination of the multiplication table confirms that every element apart from 0 has an inverse, and so this is a field.

The philosopher may sometimes love the infinite; the poet always loves the finite.

— G.K. Chesterton (1874–1936),
The Man Who Was Thursday (1908)

+	0	1	ω	ω^2
0	0	1	ω	ω^2
1	1	0	ω^2	ω
ω	ω	ω^2	0	1
ω^2	ω^2	ω	1	0

·	0	1	ω	ω^2
0	0	0	0	0
1	0	1	ω	ω^2
ω	0	ω	ω^2	1
ω^2	0	ω^2	1	ω

Table 11.1: Addition and multiplication tables for $\mathbb{Z}_2[\omega]$

How can we generalise this example to arbitrary powers of primes? The first thing to notice is that $\mathbb{Z}_2[\omega]$ is the smallest field in which the polynomial x^3-1 splits over $\mathbb{F}_2$. This is how we'll construct finite fields in general: as splitting fields of some suitably chosen polynomial over a smaller field. The next proposition identifies this smaller field:

Proposition 11.46 *A field F either has characteristic 0 and contains a subfield isomorphic to $\mathbb{Q}$, or it has prime characteristic p and contains a subfield isomorphic to $\mathbb{F}_p$.*

Proof If $\text{char}(F) = 0$ then by Proposition 9.7 F contains a subring $R = \langle 1 \rangle$ isomorphic to $\mathbb{Z}$, whose field of quotients $Q(R) \cong Q(\mathbb{Z}) = \mathbb{Q}$ is also contained in F .

If $\text{char}(F) = n$ is nonzero then F contains a subring isomorphic to $\mathbb{Z}_n$. But in this case if n is composite then there exist nonzero elements $a, b \in R \subseteq F$ such that $ab = 0$, which can't happen because F is a field and has no zero divisors. Therefore n must be equal to some prime p , and hence F contains a subring isomorphic to $\mathbb{Z}_p = \mathbb{F}_p$, which is a field. $\square$

The subfield constructed in this proof, either $Q(\langle 1 \rangle) \cong \mathbb{Q}$ if $\text{char}(F) = 0$, or $\langle 1 \rangle \cong \mathbb{F}_p$ if $\text{char}(F) = p$, is called the **prime subfield** of F . At the moment, we're interested in finite fields, whose prime subfields must therefore be of prime order. The next proposition shows that finite fields must have prime or prime power order.

Proposition 11.47 *Let F be a finite field. Then $|F| = p^n$ for some prime p and positive integer n .*

Proof By Proposition 11.46, $\text{char}(F) = p$ for some prime p , and its prime subfield $P \cong \mathbb{F}_p$. Then $F : P$ is a field extension, and since F is finite this must have finite degree $[F : P] = n$. Furthermore, F can be considered as an n -dimensional vector space over P , so we can find a basis $\{x_1, \dots, x_n\}$ such that any element $a \in F$ can be written uniquely as a P -linear combination

$$a = p_1x_1 + \dots + p_nx_n$$

where $p_1, \dots, p_n \in P$. There are p^n possible distinct P -linear combinations of this form, and hence $|F| = p^n$. $\square$

So now we know that any finite field has order p^n and a prime subfield isomorphic to $\mathbb{F}_p$. Our key example $\mathbb{Z}_2[\omega]$ has characteristic 2 and a prime subfield $\{0, 1\}$ isomorphic to $\mathbb{F}_2$. Furthermore, it consists exactly of the roots of x^3-1 together with 0. We can take this slightly further: $\mathbb{Z}_2[\omega]$ consists of all four roots of the polynomial $x^4-x = x(x-1)(x-\omega)(x-\omega^2)$ in $\mathbb{F}_2[x]$. More generally, we have the following:

Proposition 11.48 *Let F be a finite field of prime characteristic p and order $|F| = q = p^n$ for some $n \in \mathbb{N}$. Then F is a splitting field of the polynomial $f = x^q - x$ over its prime subfield $P \cong \mathbb{F}_p$, and every element of F is a root of f .*

Proof The group of units $U(F)$ consists of all invertible elements of F , and by the definition of a field this is everything but 0. So $|U(F)| = p^n - 1$. Let $a \in U(F)$. By Proposition 2.34 the order $|a|$ must divide $|U(F)|$, hence $a^{p^n-1} = 1$, so $a^{p^n} = a$, and therefore $a^{p^n} - a = 0$. Also, trivially, $0^{p^n} - 0 = 0$, so every element of F is a root of $x^{p^n} - x$. This polynomial $f = x^{p^n} - x$ splits over F , since $f = \prod_{a \in F} (x - a)$. Moreover, f cannot split over any proper subfield of F , so F must be the splitting field of f over $P \cong \mathbb{F}_p$. $\square$

We're gradually getting closer to a complete classification of finite fields: they all have prime power order and can be constructed as splitting fields of the polynomial $x^{p^n} - x$ over $\mathbb{F}_p$. All that remains is to ask how many different fields there are of a given order, up to isomorphism, and it turns out the answer is just one. To show this, we need to introduce the following function, which should look familiar.

Definition 11.49 Given a polynomial ring $R[x]$, the **formal derivative** $d: R[x] \rightarrow R[x]$ is the function that maps a polynomial

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

to the polynomial

$$df = na_n x^{n-1} + (n-1)a_{n-1} x^{n-2} + \cdots + 2a_2 x + a_1.$$

This is not a ring homomorphism, but satisfies the usual linearity and product rules for derivatives.

We can use this to give the following test for distinct roots:

Proposition 11.50 *Let F be a field, and suppose that $f \in F[x]$. Let L be a splitting field for f over F . Then the roots of f in L are distinct if and only if f and df have no non-constant common factor.*

Proof Suppose that f has a repeated root $a \in L$, so that $f = (x - a)^m g$ for some $m \geq 2$ and $g \in F[x]$. Then $df = m(x - a)^{m-1}g + (x - a)^m dg$, so f and df have the common factor $(x - a)^{m-1}$.

Conversely, suppose that f has no repeated roots. Then for each root a of f in L we have $f = (x - a)g$ where $g(a) \neq 0$. Then $df = g + (x - a)dg$ and so $df(a) = g(a) \neq 0$. By the Remainder Theorem,¹² $(x - a)$ isn't a factor of f . This is true for every root a of f in L , and hence for every factor $(x - a)$ of f in $L[x]$, so f and df have no non-constant common factors. $\square$

¹² Theorem 10.7, page 382.

Proposition 11.51 *Let p be prime, and $n \in \mathbb{N}$. Then there exists a field $\mathbb{F}_q$ of order $q = p^n$ and characteristic p , and this field is unique up to isomorphism.*

Proof Let L be the splitting field of $f = x^{p^n} - x$ over $\mathbb{F}_p$. This field has characteristic p , and hence the formal derivative $df = p^n x^{p^n-1} - 1 = -1$, which has no non-constant factors in common with f . So, by Proposition 11.50, f has no repeated roots.

Let K be the set consisting of these distinct roots; there are $\deg(f) = p^n$ of them. We claim that K is a subfield of L . Certainly it contains 0 and 1, since $0^{p^n} - 0 = 0$ and $1^{p^n} - 1 = 0$. Also, for any $a, b \in K$ we have

$$(a + b)^{p^n} = a^{p^n} + b^{p^n} = a + b$$

¹³ Proposition 8.20, page 326.

by the Freshman's Binomial Theorem,¹³ and

$$(ab)^{p^n} = a^{p^n} b^{p^n} = ab$$

and $(a^{-1})^{p^n} = (a^{p^n})^{-1} = a^{-1},$

so K is closed under addition, multiplication and inversion. This subfield K is itself the splitting field of f over $\mathbb{F}_p$: it contains all the roots of f , but no proper subfield of K does.

Hence, for any prime p and positive integer n there exists a field of order p^n with characteristic p . By Proposition 11.48, this field is the splitting field of $f = x^{p^n} - x$ over $\mathbb{F}_p$, and by Proposition 11.43 any other such field is isomorphic to K . $\square$

Definition 11.52 For any prime p and positive integer n , the finite field of characteristic p and order $q = p^n$ discussed above is the **Galois field** or **finite field** of order q . We denote it $\mathbb{F}_q$ or $\text{GF}(q)$.

I have no doubt that an author never harms his readers more than when he hides a difficulty.

— Évariste Galois (1811–1832),
Deux mémoires d'Analyse pure (1831)

11.4 Field automorphisms and the Galois group

BY THIS POINT it should be clear that there is a strong connection between solving polynomial equations and studying field extensions. In this section we will work out the details of this connection, using techniques pioneered by the French mathematician Évariste Galois (1811–1832) in the few short years before his untimely and tragic death in a duel at the age of 20.

The approach we'll take is one that has proved very fruitful on many occasions before: we will use group theory to study the symmetry of the objects under investigation, which in this case are field extensions. We must first decide what the correct notion of symmetry is.

The answer is given by our earlier discussion of simple algebraic field extensions. Proposition 11.35 in particular says that if α and β have the same minimal polynomial over some field F , then the extensions $F(\alpha):F$ and $F(\beta):F$ are F -isomorphic. When studying the internal structure of a field extension $E:F$, we want to see how many ways we can get E by extending F . Equivalently, we want to find all the different ways of mapping some field E to itself that both respects the field structure of E and keeps the embedded copy of F fixed. The concept we're heading inexorably towards is that of an F -**automorphism**: an F -isomorphism from E to itself.

It is straightforward to show that the automorphisms of a field E form a group $\text{Aut}(E)$, and it is almost as straightforward to show that the F -automorphisms of E also form a group, a subgroup of $\text{Aut}(E)$.

Definition 11.53 Let $E:F$ be a field extension. The **Galois group** $\text{Gal}(E:F)$ is the group of F -automorphisms of E ; it is a subgroup of the full automorphism group $\text{Aut}(E)$.

Let's look at a few examples.

Example 11.54 The Galois group $\text{Gal}(\mathbb{C}:\mathbb{R}) \cong \mathbb{Z}_2$. To see this, consider an arbitrary $\mathbb{R}$ -automorphism ϕ of $\mathbb{C}$. For any element $a+bi \in \mathbb{C}$ we have

$$\phi(a+bi) = \phi(a) + \phi(b)\phi(i) = a + b\phi(i)$$

so ϕ is completely determined by where it sends i . Furthermore,

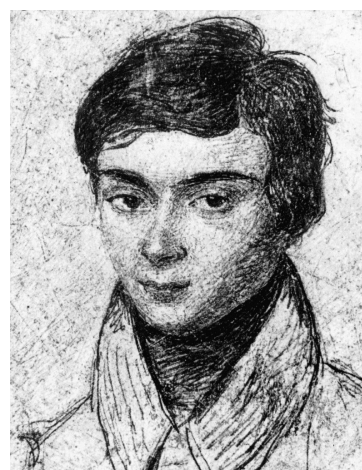
$$\phi(i)^2 = \phi(i^2) = \phi(-1) = -1$$

so $\phi(i) = \pm i$ and there are hence just two possible $\mathbb{R}$ -automorphisms of $\mathbb{C}$. If $\phi(i) = i$ then this is the identity automorphism, while if $\phi(i) = -i$ then ϕ is the conjugation map. Therefore $\text{Gal}(\mathbb{C}:\mathbb{R}) \cong \mathbb{Z}_2$.

Example 11.55 Consider the simple algebraic extension $\mathbb{Q}(\sqrt{2}):\mathbb{Q}$. A typical element of $\mathbb{Q}(\sqrt{2})$ is of the form $a+b\sqrt{2}$, for some $a, b \in \mathbb{Q}$, and a $\mathbb{Q}$ -automorphism ϕ will map this to $a+b\phi(\sqrt{2})$. So, as with the previous example, ϕ is determined completely by what it does to the adjoined element $\sqrt{2}$.

Again, we have two choices: either $\phi(\sqrt{2}) = \sqrt{2}$, in which case ϕ is the identity map, or $\phi(\sqrt{2}) = -\sqrt{2}$, a kind of $\mathbb{Q}(\sqrt{2})$ analogue of the complex conjugation map. Hence $\text{Gal}(\mathbb{Q}(\sqrt{2}):\mathbb{Q}) \cong \mathbb{Z}_2$.

Example 11.56 Let $\alpha = \sqrt[3]{5}$, and consider some $\mathbb{Q}$ -automorphism $\phi \in \text{Gal}(\mathbb{Q}(\alpha):\mathbb{Q})$. Then $\phi(\alpha)^3 = \phi(5) = 5$, so ϕ must map α to a real cube root of 5, of which α is the only one. So $\phi = \text{id}$ and hence $\text{Gal}(\mathbb{Q}(\alpha):\mathbb{Q})$ is trivial.



Wikimedia Commons

Évariste Galois (1811–1832)

Example 11.57 Any $\mathbb{Q}$ -automorphism ϕ of $\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}$ is determined by $\phi(\sqrt{2})$ and $\phi(\sqrt{3})$. The existence of two adjoined elements raises the possibility that ϕ might map one to the other but, in this case at least, this can't happen. The reason for this is that

$$\phi(\sqrt{2})^2 = \phi(\sqrt{2}^2) = \phi(2) = 2$$

$$\text{and } \phi(\sqrt{3})^2 = \phi(\sqrt{3}^2) = \phi(3) = 3$$

so ϕ can only map $\sqrt{2}$ to $\pm\sqrt{2}$ and $\sqrt{3}$ to $\pm\sqrt{3}$. There are therefore four different $\mathbb{Q}$ -automorphisms of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$:

ϕ	$\phi(\sqrt{2})$	$\phi(\sqrt{3})$
$\phi_1 = \text{id}$	$\sqrt{2}$	$\sqrt{3}$
ϕ_2	$\sqrt{2}$	$-\sqrt{3}$
ϕ_3	$-\sqrt{2}$	$\sqrt{3}$
ϕ_4	$-\sqrt{2}$	$-\sqrt{3}$

Each of these has order 2, so $\text{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}) \cong V_4 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

In Example 11.54 we saw that the only nontrivial $\mathbb{R}$ -automorphism in $\mathbb{C}$ is the complex conjugation map swapping i and $-i$. Something very similar happened in Example 11.55, where there are again two $\mathbb{Q}$ -automorphisms: the identity, and the one that swaps $\sqrt{2}$ and $-\sqrt{2}$. In Example 11.57 things are slightly more complicated, with a total of four $\mathbb{Q}$ -automorphisms.

Motivated by this, and the other examples above, we introduce the following definition:

Definition 11.58 Let $E : F$ be an algebraic extension. Then $\alpha, \beta \in E$ are **F -conjugate** if there is an F -automorphism of E mapping α to β . Equivalently α and β are F -conjugate if they have the same minimal polynomial over F .

The equivalence of the two criteria in the above definition is a consequence of the following:

Proposition 11.59 Let $E : F$ be a field extension, let $f \in F[x]$ be a polynomial over F , and let $\phi \in \text{Gal}(E : F)$ be an F -automorphism of E . Then $\phi(f(\alpha)) = f(\phi(\alpha))$ for any $\alpha \in E$, and in particular ϕ permutes the roots of f in E .

Proof Let $f = a_n x^n + \cdots + a_1 x + a_0$ for some $a_0, \dots, a_n \in F$. Then

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

for any $\alpha \in E$. In particular, if α is a root of f then $\phi(\alpha)$ is too, since $f(\phi(\alpha)) = \phi(f(\alpha)) = \phi(0) = 0$. Since ϕ is a field automorphism, it is a bijection from E to itself, so it permutes the roots of f amongst themselves. $\square$

Although understanding field extensions is a worthwhile goal in its own right, a major part of our motivation for doing so is to study the solubility of polynomial equations. Proposition 11.59 presents a link between the roots of polynomials in $F[x]$ and the F -automorphisms of extensions containing those roots. We can think of the Galois group $\text{Gal}(E:F)$ as a group of symmetries, not just of the field extension $E:F$ but also of the polynomials over F whose roots are contained in E . To that end, we introduce the following definition, which extends Definition 11.53:

Definition 11.60 Let $f \in F[x]$ be a polynomial, and let E be the splitting field of f over F . Then the **Galois group** $\text{Gal}(f)$ or $\text{Gal}(f, F)$ of f over F is the group $\text{Gal}(E:F)$ of the splitting field extension of f over F .

Calculating Galois groups is often tricky, but we will look at some examples now.

Example 11.61 What is $\text{Gal}(x^3 - 1, \mathbb{Q})$? We need to find the splitting field of $f = x^3 - 1$ over $\mathbb{Q}$, and first we notice that this polynomial factorises as $(x - 1)(x^2 + x + 1)$. The second factor $g = x^2 + x + 1$ is irreducible over $\mathbb{Q}$, so the splitting field of f must be the smallest field in which g splits. This is $\mathbb{Q}(\omega)$ where $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. Any $\mathbb{Q}$ -automorphism ϕ of $\mathbb{Q}(\omega)$ must map ω to either ω or ω^2 , since $\phi(\omega)^3 = \phi(1) = 1$, but we can't have $\phi(\omega) = 1$ because that isn't an automorphism. There are thus two possibilities: $\phi_1 = \text{id}$ and ϕ_2 given by $\phi_2(\omega) = \omega^2$. Hence $|\text{Gal}(x^3 - 1, \mathbb{Q})| = 2$, and the only possibility is $\text{Gal}(x^3 - 1, \mathbb{Q}) \cong \mathbb{Z}_2$.

The next example is a slight variation with markedly different results.

Example 11.62 In Example 11.41 we investigated the splitting field of the polynomial $f = x^3 - 5$ over $\mathbb{Q}$. Let's now calculate the Galois group $\text{Gal}(f, \mathbb{Q})$ of this polynomial. Recall from the earlier discussion that f has three roots: α , $\alpha\omega$ and $\alpha\omega^2$, where $\alpha = \sqrt[3]{5}$ and $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$.

Any $\mathbb{Q}$ -automorphism ϕ of the splitting field $\mathbb{Q}(\alpha, \omega)$ is determined by what it does to α and ω . Moreover, we require $\phi(\alpha)^3 = \phi(5) = 5$ and $\phi(\omega)^3 = \phi(1) = 1$. There are three possibilities for α , namely $\phi(\alpha) = \alpha$, $\alpha\omega$ or $\alpha\omega^2$, and two possibilities for ω , namely $\phi(\omega) = \omega$ or ω^2 . (We can't have $\phi(\omega) = 1$ because that wouldn't be an

ϕ	$\phi(\alpha)$	$\phi(\omega)$
$\phi_1 = \text{id}$	α	ω
ϕ_2	$\alpha\omega$	ω
ϕ_3	$\alpha\omega^2$	ω
ϕ_4	α	ω^2
ϕ_5	$\alpha\omega$	ω^2
ϕ_6	$\alpha\omega^2$	ω^2

Table 11.2: Elements of $\text{Gal}(x^3 - 5, \mathbb{Q})$

ϕ	$\phi(\alpha)$	$\phi(\alpha\omega)$	$\phi(\alpha\omega^2)$
$\phi_1 = \text{id}$	α	$\alpha\omega$	$\alpha\omega^2$
ϕ_2	$\alpha\omega$	$\alpha\omega^2$	α
ϕ_3	$\alpha\omega^2$	α	$\alpha\omega$
ϕ_4	α	$\alpha\omega^2$	$\alpha\omega$
ϕ_5	$\alpha\omega$	α	$\alpha\omega^2$
ϕ_6	$\alpha\omega^2$	$\alpha\omega$	α

Table 11.3: Permutations of the roots of $x^3 - 5$

automorphism.) So we have six possibilities, listed in Table 11.2.

All of these are $\mathbb{Q}$ -automorphisms of $\mathbb{Q}(\alpha, \omega)$ so $|\text{Gal}(x^3 - 5, \mathbb{Q})| = 6$. We thus have two possibilities for the Galois group: either $\mathbb{Z}_6$ or S_3 . We can figure out which one it is in various ways. We could, for example, consider the orders of the elements:

$$|\phi_1| = 1, \quad |\phi_2| = |\phi_3| = 3, \quad |\phi_4| = |\phi_5| = |\phi_6| = 2.$$

There is no element of order 6, so this can't be $\mathbb{Z}_6$ and must therefore be S_3 instead. We can also see this from the way the automorphisms permute the roots of f , shown in Table 11.3. Each of the six possible permutations is represented, so $\text{Gal}(x^3 - 5, \mathbb{Q}) \cong S_3$.

The following example is rather more complicated.

Example 11.63 Let $f = x^4 - 5x^2 + 5$. We want to calculate $\text{Gal}(f, \mathbb{Q})$, and to do this we start by finding the roots of f . In general, finding the roots of a quartic polynomial is somewhat involved (the formula derived by Lodovico Ferrari in the sixteenth century is rather complicated) but in this case we can view it as a quadratic polynomial in x^2 , and solve it in the usual way to get

$$x^2 = \frac{5}{2} \pm \frac{\sqrt{5}}{2}.$$

The roots of f are therefore

$$\begin{aligned} \alpha_1 &= \sqrt{\frac{5}{2} + \frac{\sqrt{5}}{2}}, & \alpha_2 &= \sqrt{\frac{5}{2} - \frac{\sqrt{5}}{2}}, \\ \alpha_3 &= -\sqrt{\frac{5}{2} + \frac{\sqrt{5}}{2}} = -\alpha_1, & \alpha_4 &= -\sqrt{\frac{5}{2} - \frac{\sqrt{5}}{2}} = -\alpha_2. \end{aligned}$$

The Galois group $G = \text{Gal}(f, \mathbb{Q})$ consists of all $\mathbb{Q}$ -automorphisms of the splitting field of f , which is $\mathbb{Q}(\alpha_1, \alpha_2)$. These permute the roots amongst themselves, and hence $G \leq S_4$.

Is $\mathbb{Q}(\alpha_1, \alpha_2)$ the simplest way of writing this splitting field? Actually we can simplify it further. Observe that $\alpha_1\alpha_2 = \sqrt{5}$, and also that $2\alpha_1^2 = 5 + \sqrt{5}$. Hence $\alpha_1\alpha_2 = 2\alpha_1^2 - 5$, and so $\alpha_2 = 2\alpha_1 - 5/\alpha_1$. Furthermore, $\alpha_3 = -\alpha_1$ and $\alpha_4 = -\alpha_2$, so we can recover all the other roots from α_1 and the required splitting field is $\mathbb{Q}(\alpha_1)$. The polynomial f is irreducible (by Eisenstein's Criterion) and monic, and is hence the minimal polynomial for the extension $\mathbb{Q}(\alpha_1) : \mathbb{Q}$, which must therefore have degree 4.

Suppose that $\phi \in G$, and that $\phi(\alpha_1) = \alpha_2$. Then

$$\phi(\sqrt{5}) = 2\phi(\alpha_1)^2 - 5 = 2\alpha_2^2 - 5 = -\sqrt{5},$$

and hence

$$\phi(\alpha_2) = \phi(\sqrt{5}/\alpha_1) = -\sqrt{5}/\alpha_2 = -\alpha_1 = \alpha_3.$$

Furthermore,

$$\phi(\alpha_3) = \phi(-\alpha_1) = -\alpha_2 = \alpha_4, \quad \phi(\alpha_4) = \phi(-\alpha_2) = -\alpha_3 = \alpha_1.$$

So ϕ , and for that matter any other $\mathbb{Q}$ -automorphism that maps α_1 to α_2 , cyclicly permutes the four roots. Therefore $\langle \phi \rangle \cong \mathbb{Z}_4$. Any $\mathbb{Q}$ -automorphism mapping α_1 to α_4 must similarly be equal to ϕ^{-1} .

Now consider a $\mathbb{Q}$ -automorphism ψ that maps α_1 to $\alpha_3 = -\alpha_1$. Then

$$\psi(\sqrt{5}) = 2\psi(\alpha_1)^2 - 5 = 2\alpha_1^2 - 5 = \sqrt{5}$$

and so

$$\psi(\alpha_2) = \psi(\sqrt{5}/\alpha_1) = -\sqrt{5}/\alpha_1 = -\alpha_2 = \alpha_4.$$

Furthermore, $\psi(\alpha_3) = \alpha_1$ and $\psi(\alpha_4) = \alpha_2$, so $\psi = \phi^2$.

These are the only possibilities, and so $G \cong \langle \phi \rangle \cong \mathbb{Z}_4$.

11.5 The Galois Correspondence

IT'S ILLUMINATING to look at subgroups of the Galois group of a field extension $E:F$, and see how the various F -automorphisms act on the various subfields of E . In Example 11.57 we saw that

$$\text{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3}):\mathbb{Q}) = \{\phi_1 = \text{id}, \phi_2, \phi_3, \phi_4\} \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2.$$

Every element of this group fixes the base field $\mathbb{Q}$. Every element of the subgroup $\{\phi_1, \phi_2\}$ fixes the subfield $\mathbb{Q}(\sqrt{2})$, but ϕ_2 acts nontrivially on $\mathbb{Q}(\sqrt{3})$ so the subgroup doesn't fix this subfield. Similarly, $\{\phi_1, \phi_3\}$ leaves $\mathbb{Q}(\sqrt{3})$ invariant but not $\mathbb{Q}(\sqrt{2})$. Neither $\mathbb{Q}(\sqrt{2})$ nor $\mathbb{Q}(\sqrt{3})$ are fixed by $\{\phi_1, \phi_4\}$, but $\sqrt{6} = \sqrt{2}\sqrt{3}$ is mapped to itself by both ϕ_1 and ϕ_4 , so the subfield $\mathbb{Q}(\sqrt{6})$ is fixed by this subgroup. Finally, the trivial subgroup $\{\phi_1\}$ leaves $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ invariant.

Figure 11.1 shows the subgroup lattice for $\text{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3}):\mathbb{Q})$, with inclusions going up the page. Figure 11.2 shows the subfield lattice for $\mathbb{Q}(\sqrt{2}, \sqrt{3})$, this time drawn upside down, with inclusions going down the page. We'll come to the reason for this reversal soon, but for the moment observe the striking and suggestive similarity between the two diagrams. Each subfield in Figure 11.2 is fixed by the corresponding subgroup in Figure 11.1. This is an example of the **Galois correspondence**, which is the subject of the rest of this section.

Let's try to generalise this construction. Given an extension $E:F$, we want to define a correspondence between groups and fields, so that on one hand we have subgroups of the Galois group $\text{Gal}(E:F)$, and on the other we have subfields K of E , such that $F \subseteq K \subseteq E$. The latter is something we'll run into a lot over the rest of this chapter, so it would be a good idea to introduce some appropriate terminology.

It's people like that who make you realise how little you've accomplished. It is a sobering thought, for example, that when Mozart was my age, he had been dead for two years.

— Tom Lehrer,
introduction to *Alma*,
from: *That Was The Year That Was*
(1965)

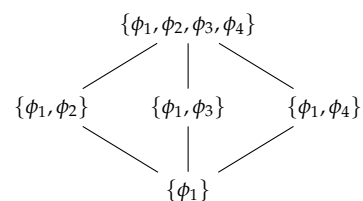


Figure 11.1: Subgroup lattice diagram for $\text{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3}):\mathbb{Q}) \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$

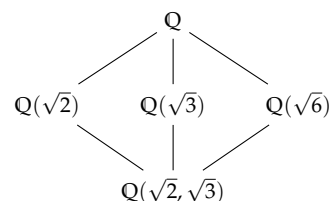


Figure 11.2: Subfield lattice diagram for $\mathbb{Q}(\sqrt{2}, \sqrt{3})$

Definition 11.64 An **intermediate subfield** of an extension $E:F$ is a subfield K of E that contains F , such that $F \subseteq K \subseteq E$.

Firstly, we want to define a map from subgroups of $\text{Gal}(E:F)$ to intermediate subfields of $E:F$. In our key example above, we identified each subgroup with the subfield that it left invariant. So, for any subgroup $H \leq \text{Gal}(E:F)$ we define the **fixed field** of H to be¹⁴

¹⁴ This is sometimes denoted E^H .

$$\mathcal{F}(H) = \text{Fix}_H(E) = \{a \in E : \phi(a) = a \text{ for all } \phi \in H\}.$$

Secondly, going back the other way, we set

$$\mathcal{G}(K) = \text{Gal}(E:K) = \{\phi \in \text{Gal}(E:F) : \phi(k) = k \text{ for all } k \in K\}$$

for any intermediate subfield K .

It would all be rather neat and tidy if $\mathcal{F}$ and $\mathcal{G}$ were inverses of each other. But this isn't always the case, as the next two examples show.

Example 11.65 We saw in Example 11.56 that the Galois group of the extension $\mathbb{Q}(\sqrt[3]{5}):\mathbb{Q}$ is trivial. Therefore $\mathcal{G}(\mathbb{Q})$ must also be trivial. In particular, $\mathcal{F}(\mathcal{G}(\mathbb{Q})) = \mathbb{Q}(\sqrt[3]{5})$. So $\mathcal{F}$ is not the inverse of $\mathcal{G}$ in this case.

The problem with this example, or at least the thing that stops $\mathcal{F}$ and $\mathcal{G}$ being inverses of each other, is that the minimal polynomial of $\alpha = \sqrt[3]{5}$ is $x^3 - 5$. But the splitting field of this isn't $\mathbb{Q}(\alpha)$, it's $\mathbb{Q}(\alpha, \omega)$, where $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. The next example also fails, but for a different reason.

Example 11.66 Let F be a field, and adjoin some element $\alpha \notin F$ to get $F(\alpha)$. The polynomial ring $F(\alpha)[x]$ consists, as usual, of all finite-degree polynomials with coefficients in $F(\alpha)$. In particular, $x^n - \alpha$ is irreducible over $F(\alpha)$ by Eisenstein's Criterion. Now let $\beta = \alpha^{1/n}$ be the n th root of α . Then $F(\alpha, \beta) = F(\alpha, \alpha^{1/n}) = F(\alpha^{1/n}) = F(\beta)$ is a simple extension of F with degree n .

Now consider the case where $F = \mathbb{F}_p$ for some prime p . Then

$$(x^p - \alpha) = (x - \alpha^{1/p})^p = (x - \beta)^p,$$

by the Freshman's Binomial Theorem,¹⁵ and so $x^p - \alpha$ has only one root in $\mathbb{F}_p(\beta)$, and indeed in any other extension of F . Therefore $\text{Gal}(\mathbb{F}_p(\beta):\mathbb{F}_p(\alpha))$ is trivial, and hence $\mathcal{G}(\mathbb{F}_p(\alpha))$, the group of all $\mathbb{F}_p(\alpha)$ -automorphisms of $\mathbb{F}_p(\beta)$, must be trivial. Applying $\mathcal{F}$ to this we get the field consisting of every element of $\mathbb{F}_p(\beta)$ fixed by the identity $\mathbb{F}_p(\alpha)$ -automorphism, which is all of $\mathbb{F}_p(\beta)$. So $\mathcal{F}(\mathcal{G}(\mathbb{F}_p(\alpha))) = \mathbb{F}_p(\beta)$ and hence $\mathcal{F}$ and $\mathcal{G}$ aren't inverses of each other.

¹⁵ Proposition 8.20, page 326.

The problem here is that $x^p - \alpha$ only has a single root in the splitting field $\mathbb{F}_p(\beta)$; moreover we can adjoin as many other elements to $\mathbb{F}_p(\beta)$

as we like, and it will still only have a single root, because $x^p - \alpha = (x - \beta)^p$ over $\mathbb{F}_p$. We have a single root, repeated p times.

We want to find a well-defined characterisation of nice field extensions or polynomials where neither of these awkward situations arise. We want to know what criteria an extension $E:F$ must satisfy to ensure $\mathcal{F}$ and $\mathcal{G}$ are inverses of each other, so that there is a neat and tidy correspondence between intermediate subfields of $E:F$ and subgroups of $\text{Gal}(E:F)$.

Before we get to that, we have a few useful facts about $\mathcal{F}$ and $\mathcal{G}$ to prove. The first of these says that while $\mathcal{F}$ and $\mathcal{G}$ aren't necessarily inverses, we do at least have $K \subseteq \mathcal{F}(\mathcal{G}(K))$ and $H \leq \mathcal{G}(\mathcal{F}(H))$ for any intermediate subfield K and subgroup H .

Proposition 11.67 *Let K be an intermediate subfield of an extension $E:F$ and suppose $H \leq \text{Gal}(E:F)$. Then $K \subseteq \mathcal{F}(\mathcal{G}(K))$ and $H \leq \mathcal{G}(\mathcal{F}(H))$.*

Proof The group $\mathcal{G}(K)$ comprises all of the F -automorphisms in $\text{Gal}(E:F)$ that keep K fixed pointwise. Any element $k \in K$ is thus fixed by everything in $\mathcal{G}(K) = \text{Gal}(E:K)$ and hence $k \in \text{Fix}_{\mathcal{G}(K)}(E) = \mathcal{F}(\mathcal{G}(K))$. Thus $K \subseteq \mathcal{F}(\mathcal{G}(K))$.

Now considering a subgroup $H \leq \text{Gal}(E:F)$, the field $\mathcal{F}(H) = \text{Fix}_H(E)$ consists of every element of E fixed by the F -automorphisms in H . So any $h \in H$ fixes everything in $\mathcal{F}(H)$, hence $h \in \mathcal{G}(\mathcal{F}(H))$ and therefore $H \leq \mathcal{G}(\mathcal{F}(H))$. $\square$

We've seen so far that there is at least a partial correspondence between intermediate fields and subgroups of the Galois group. In ideal circumstances, the exact details of which we will shortly investigate, this correspondence is bijective, but as we saw above there are awkward cases in which it isn't. In general, however, there is at least some link between a given subgroup H of the Galois group $\text{Gal}(E:F)$ and the extension $E:\mathcal{F}(H)$:

Proposition 11.68 *Let $E:F$ be a finite field extension, and let $H \leq \text{Aut}(E)$ be a finite subgroup of the automorphism group of E . Then*

$$[E:\mathcal{F}(H)] = |H|.$$

To prove this we need a technical lemma concerning linear independence of field monomorphisms, due to Richard Dedekind (1831–1916):

Lemma 11.69 (Richard Dedekind) *Let F_1 and F_2 be fields and suppose that $f_1, \dots, f_n: F_1 \rightarrow F_2$ are distinct monomorphisms. Then $f_1, \dots, f_n$ are linearly independent.*

Proof We prove this lemma by induction and contradiction. The result is obviously true for $n = 1$, since the only coefficient $a_1 \in F_2$ for

which $a_1 f_1 = 0$ is $a_1 = 0$.

Now suppose that the result is true for all $n < k$, for some arbitrary $k \in \mathbb{N}$. That is, every set of fewer than k distinct monomorphisms is linearly independent. Assume that the result fails for $n = k$; that is, there exist k distinct monomorphisms $f_1, \dots, f_k: F_1 \rightarrow F_2$ and coefficients $a_1, \dots, a_k \in F_2$, not all of which are zero, such that

$$a_1 f_1(x) + \dots + a_k f_k(x) = 0 \quad (11.3)$$

for all $x \in F_1$. In fact, we may assume that all of $a_1, \dots, a_k$ are nonzero, otherwise we have a linearly dependent set of fewer than k distinct monomorphisms, contrary to the inductive hypothesis.

Since the monomorphisms are distinct, and in particular $f_1 \neq f_k$, there exists some nonzero element $y \in F_1$ such that $f_1(y) \neq f_k(y)$. Then

$$a_1 f_1(xy) + \dots + a_k f_k(xy) = 0$$

and so

$$a_1 f_1(x) f_1(y) + \dots + a_k f_k(x) f_k(y) = 0. \quad (11.4)$$

Multiplying (11.3) by $f_1(y)$ and subtracting (11.4) gives

$$a_2(f_1(y) - f_2(y))f_2(x) + \dots + a_k(f_1(y) - f_k(y))f_k(x) = 0. \quad (11.5)$$

This is an F_2 -linear combination of fewer than k distinct monomorphisms $f_2, \dots, f_k: F_1 \rightarrow F_2$. The coefficient of $f_i(x)$ in this equation is $a_i(f_1(y) - f_i(y))$, for $2 \leq i \leq k$. All of $a_2, \dots, a_k$ are nonzero, and at the very least the coefficient $a_k(f_1(y) - f_k(y))$ must be nonzero, since we have chosen y such that $f_1(y) \neq f_k(y)$. Thus (11.5) implies that $f_2, \dots, f_k$ are linearly independent, contradicting the inductive hypothesis.

Therefore, no equation of the form (11.3) exists, and hence $f_1, \dots, f_k$ are linearly independent.

By induction, any finite set of monomorphisms $f_1, \dots, f_n$ is linearly independent, as claimed. $\square$

We can now prove Proposition 11.68:

Proof of Proposition 11.68 Let $|H| = n$ and let $[E: \mathcal{F}(H)] = m$. Suppose that $H = \{h_1 = e, \dots, h_n\}$ and that $S = \{x_1, \dots, x_m\}$ is a basis for E considered as an m -dimensional vector space over $\mathcal{F}(H)$. E is therefore isomorphic to $\mathcal{F}(H)^m$.

We prove the equality by contradiction, showing that m can be neither strictly greater than, nor strictly less than n .

First suppose that $[E: \mathcal{F}(H)] = m > n = |H|$, and consider the $n \times m$ matrix

$$A = \begin{bmatrix} h_1(x_1) & \dots & h_m(x_1) \\ \vdots & & \vdots \\ h_1(x_n) & \dots & h_m(x_n) \end{bmatrix}.$$

We can view this as representing a linear map $\mathcal{F}(H)^m \rightarrow \mathcal{F}(H)^n$. By the Dimension Theorem we have

$$\text{rank}(A) + \text{nullity}(A) = n.$$

The rows of A are linearly independent, so $\text{rank}(A) = n$ and hence $\text{nullity}(A) = m - n \geq 1$. Hence $\ker(A)$ is nontrivial, which means that there exist $y_1, \dots, y_m \in \mathcal{F}(H)^n$, not all zero, such that:

$$\begin{aligned} h_1(x_1)y_1 + \cdots + h_m(x_1)y_m &= 0 \\ &\vdots \\ h_1(x_n)y_1 + \cdots + h_m(x_n)y_m &= 0 \end{aligned} \quad (11.6)$$

We can write any element $a \in E$ as a unique $\mathcal{F}(H)$ -linear combination

$$a = a_1x_1 + \cdots + a_mx_m. \quad (11.7)$$

Multiplying the equations in (11.6) by the coefficients $a_1, \dots, a_m$ respectively, we get:

$$\begin{aligned} a_1h_1(x_1)y_1 + \cdots + a_1h_n(x_1)y_n &= 0 \\ &\vdots \\ a_mh_1(x_m)y_1 + \cdots + a_mh_n(x_m)y_n &= 0 \end{aligned} \quad (11.8)$$

All of the coefficients $a_1, \dots, a_m$ lie in $\mathcal{F}(H)$ and are hence fixed by all of the $\mathcal{F}(H)$ -automorphisms $h_1, \dots, h_n \in H$. What this means is that $h_i(a_j) = a_j$ for all i and j , and so we can rewrite (11.7) as

$$\begin{aligned} h_1(a_1)h_1(x_1)y_1 + \cdots + h_n(a_1)h_n(x_1)y_n &= 0 \\ &\vdots \\ h_1(a_m)h_1(x_m)y_1 + \cdots + h_n(a_m)h_n(x_m)y_n &= 0 \end{aligned} \quad (11.9)$$

and then as

$$\begin{aligned} h_1(a_1x_1)y_1 + \cdots + h_n(a_1x_1)y_n &= 0 \\ &\vdots \\ h_1(a_mx_m)y_1 + \cdots + h_n(a_mx_m)y_n &= 0 \end{aligned} \quad (11.10)$$

Adding all of these equations together and using (11.7) we get

$$y_1h_1(a) + \cdots + y_nh_n(a) = 0. \quad (11.11)$$

This is true for any $a \in E$, and since the elements $y_1, \dots, y_n$ aren't all zero, this means that the automorphisms $h_1, \dots, h_n$ are linearly dependent, contrary to Lemma 11.69. Hence our original assertion, that $|H| > [E : \mathcal{F}(H)]$, can't be true, so instead $|H| \leq [E : \mathcal{F}(H)]$.

Now suppose $|H| = n < m = [E : \mathcal{F}(H)]$. Let $T = \{x_1, \dots, x_{n+1}\} \subseteq S$ be a set of $n+1$ linearly independent vectors in E regarded as an m -

dimensional vector space over $\mathcal{F}(H)$. Consider the $m \times (n+1)$ matrix

$$B = \begin{bmatrix} h_1(x_1) & \dots & h_1(x_{n+1}) \\ \vdots & & \vdots \\ h_n(x_1) & \dots & h_n(x_{n+1}) \end{bmatrix}$$

Again, we can view this as representing a linear map $\mathcal{F}(H)^{n+1} \rightarrow \mathcal{F}(H)^n$, and by the Dimension Theorem we have

$$\text{rank}(B) + \text{nullity}(B) = n + 1.$$

The rows of this matrix are linearly independent, so $\text{rank}(B) = n$ and hence $\text{nullity}(B) = 1$, which means that $\ker(B)$ is nontrivial. Hence we can find elements $z_1, \dots, z_{n+1} \in E$, not all zero, such that:

$$\begin{aligned} h_1(x_1)z_1 + \dots + h_1(x_{n+1})z_{n+1} &= 0 \\ &\vdots \end{aligned} \tag{11.12}$$

$$h_n(x_1)z_1 + \dots + h_n(x_{n+1})z_{n+1} = 0$$

Suppose that $z_1, \dots, z_{n+1}$ are chosen such that as few of them as possible are zero. Furthermore, arrange them so that $z_1, \dots, z_r$ are nonzero, and $z_{r+1} = \dots = z_{n+1} = 0$. We can thus rewrite (11.12) as

$$\begin{aligned} h_1(x_1)z_1 + \dots + h_1(x_r)z_r &= 0 \\ &\vdots \end{aligned} \tag{11.13}$$

$$h_n(x_1)z_1 + \dots + h_n(x_r)z_r = 0$$

Choose some element $h \in H$ and apply it to all of the equations in (11.13) to get

$$\begin{aligned} hh_1(x_1)h(z_1) + \dots + hh_1(x_r)h(z_r) &= 0 \\ &\vdots \end{aligned} \tag{11.14}$$

$$hh_n(x_1)h(z_1) + \dots + hh_n(x_r)h(z_r) = 0$$

Recall that $H = \{h_1, \dots, h_n\}$ is a finite group, and so multiplying every element on the left by h merely permutes the set, so $hH = \{hh_1, \dots, hh_n\} = H$. Hence we can reorder the equations in (11.14) as

$$\begin{aligned} h_1(x_1)h(z_1) + \dots + h_1(x_r)h(z_r) &= 0 \\ &\vdots \end{aligned} \tag{11.15}$$

$$h_n(x_1)h(z_1) + \dots + h_n(x_r)h(z_r) = 0$$

We now multiply (11.13) by $h_1(z_1)$ and (11.15) by z_1 , and subtract the latter from the former to get

$$\begin{aligned} (h(z_1)z_2 - h(z_2)z_1)h_1(x_2) + \dots + (h(z_1)z_r - h(z_r)z_1)h_1(x_r) &= 0 \\ &\vdots \\ (h(z_1)z_2 - h(z_2)z_1)h_n(x_2) + \dots + (h(z_1)z_r - h(z_r)z_1)h_n(x_r) &= 0 \end{aligned}$$

This is a system of n equations with $(r-1)$ terms. But we carefully chose $z_1, \dots, z_r$ such that these were the smallest possible number of nonzero elements, and have now constructed a system of equations with fewer nonzero coefficients. This is a contradiction, unless this new system of equations is trivial; that is, all of the coefficients are zero. If this is the case, then we have $h(z_1)z_i - h(z_i)z_1 = 0$ for $2 \leq i \leq r$. Rearranging this, we find that

$$h(z_1 z_i^{-1}) = z_1 z_i^{-1}.$$

But the choice of h was arbitrary, so $z_1 z_i^{-1} \in \mathcal{F}(H)$. Therefore we can find some nonzero element $c \in E$ and elements $w_1, \dots, w_r \in \mathcal{F}(H)$ such that $z_i = cw_i$ for all i .

Then the first equation in (11.13), recalling that h_1 is the identity, becomes

$$x_1 cw_1 + \dots + x_r cw_r = 0.$$

Dividing through by c gives

$$x_1 w_1 + \dots + x_r w_r = 0$$

which implies that $x_1, \dots, x_r$ are linearly dependent, contrary to our initial assumption. Therefore $|H| = n \geq m = [E : \mathcal{F}(H)]$.

Together with the first part, this implies $|H| = [E : \mathcal{F}(H)]$. □

Now we'll try to nail down the exact conditions under which the maps $\mathcal{F}$ and $\mathcal{G}$ are mutually inverse, implying a bijective correspondence between intermediate subfields and subgroups of the Galois group.

Our main clues for this investigation come from Examples 11.65 and 11.66, which provide two different ways in which this correspondence can fail to be bijective. In fact these are the only two ways the correspondence can fail, but we need to do a little more work to prove that.

First we'll consider the failure mode seen in Example 11.65. The issue there was that $x^3 - 5$ is the minimal polynomial over $\mathbb{Q}$ for $\alpha = \sqrt[3]{5}$, but it doesn't split in $\mathbb{Q}(\alpha)$. We can avoid this situation if we restrict our attention to extensions in which every polynomial splits:

Definition 11.70 A field extension $E : F$ is **normal** if and only if every irreducible polynomial in $F[x]$ that has at least one root in E splits completely over E .

In other words, $E : F$ is normal exactly when any polynomial $f \in F[x]$ with a root in E has all its roots in E .

Example 11.71 The extension $\mathbb{C} : \mathbb{R}$ is normal, since every polynomial in $\mathbb{R}[x]$ splits completely over $\mathbb{C}$.

The next example is slightly more complicated.

Example 11.72 The extension $\mathbb{Q}(\sqrt{2}) : \mathbb{Q}$ is normal. Suppose that $f \in \mathbb{Q}[x]$ is irreducible over $\mathbb{Q}$ and has a root $\alpha = a + b\sqrt{2}$ in $\mathbb{Q}(\sqrt{2})$. Then f factorises as $(x - \alpha)g$ in $\mathbb{Q}(\alpha)$. But for f to be in $\mathbb{Q}[x]$ the polynomial g must also have some factor $(x - \beta)$, where $\beta = c + d\sqrt{2}$, such that $(x - \alpha)(x - \beta) = x^2 - (\alpha + \beta)x + \alpha\beta \in \mathbb{Q}[x]$. For $\alpha + \beta \in \mathbb{Q}$ we require $d = -b$, and for $\alpha\beta \in \mathbb{Q}$ we require $a = c$. Hence $\beta = \bar{\alpha} = a - b\sqrt{2}$. So f has $h = x^2 - 2ax + (a^2 + 2b^2) \in \mathbb{Q}[x]$ as a factor. But f is irreducible over $\mathbb{Q}$, so $f = kh$ for some $k \in \mathbb{Q}$. Hence f splits completely over $\mathbb{Q}(\sqrt{2})$, and so $\mathbb{Q}(\sqrt{2}) : \mathbb{Q}$ is a normal extension.

This is a rather cumbersome way of proving normality: we have to show that any irreducible polynomial with a root in the extension field has all its roots in the extension field, which is manageable for something relatively simple like $\mathbb{Q}(\sqrt{2}) : \mathbb{Q}$ but rapidly becomes problematic for more complicated extensions. The following result simplifies matters considerably.

Proposition 11.73 *A finite extension $E : F$ is normal if and only if it is a splitting field for some polynomial in $F[x]$.*

Proof First suppose that $E : F$ is normal, and that $[E : F] = n$. Then we can view E as an n -dimensional vector space over F . Let $S = \{\alpha_1, \dots, \alpha_n\}$ be a basis for this space, and let f_i be the minimal polynomial of the adjoined element α_i , for $i = 1, \dots, n$. Since $E : F$ is normal, any polynomial in $F[x]$ with a root in E splits completely over E ; in particular, each minimal polynomial f_i splits completely over E . Let $f = f_1 f_2 \dots f_n$ be the product of all the minimal polynomials. Then f also splits completely over E , but not over any proper subfield, because E is generated by S and not by any proper subset thereof. Hence E is a splitting field for f over F .

Conversely, suppose that E is a splitting field for some polynomial $f \in F[x]$. Let $g \in F[x]$ be some arbitrary polynomial of degree at least 2, that is irreducible over F and has a root $\alpha \in E$.

Let K be the splitting field of fg over F , and suppose that β is some other root of g in K . We can construct a lattice of fields as depicted in Figure 11.3. From this diagram, using the Tower Law,¹⁶ we can deduce the following identities:

$$[E(\alpha) : E][E : F] = [E(\alpha) : F] = [E(\alpha) : F(\alpha)][F(\alpha) : F] \quad (11.16)$$

$$[E(\beta) : E][E : F] = [E(\beta) : F] = [E(\beta) : F(\beta)][F(\beta) : F] \quad (11.17)$$

By Proposition 11.35, since α and β are roots of the same irreducible

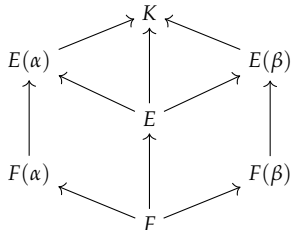


Figure 11.3: Lattice diagram for the fields in Proposition 11.73

¹⁶ Proposition 11.21, page 428.

polynomial g , there is an F -isomorphism $\phi: F(\alpha) \rightarrow F(\beta)$. Then

$$[F(\alpha):F] = [F(\beta):F].$$

Since E is a splitting field for f over F , it follows that $E(\alpha)$ and $E(\beta)$ are splitting fields for f over, respectively, $F(\alpha)$ and $F(\beta)$. Then by Proposition 11.43 there is an isomorphism $\psi: E(\alpha) \rightarrow E(\beta)$ extending ϕ . Hence

$$[E(\alpha):F(\alpha)] = [E(\beta):F(\beta)].$$

Furthermore, $[E(\alpha):E] = 1$ since $\alpha \in E$. Putting all this together with (11.16) and (11.17) we get

$$\begin{aligned} [E(\beta):E][E:F] &= [E(\beta):F(\beta)][F(\beta):F] \\ &= [E(\alpha):F(\alpha)][F(\alpha):F] \\ &= [E(\alpha):E][E:F] \\ &= [E:F]. \end{aligned}$$

Hence $[E(\beta):E] = 1$, which can only be the case if $\beta \in E$. So if E is a splitting field for some polynomial $f \in F[x]$, then any other irreducible polynomial $g \in F[x]$ that has a root in E has all of its roots in E , and hence $E:F$ is normal. $\square$

Proposition 11.74 *Let K be an intermediate subfield of a finite normal extension $E:F$. Every F -monomorphism of K into E can be extended to a F -automorphism of E .*

Proof Let $\phi: K \rightarrow E$ be an F -monomorphism. By Proposition 11.73 there is a polynomial $f \in F[x]$ with splitting field E . Since $F \subseteq K$ and $F \subseteq \phi(K)$, it follows that $F[x] \subseteq K[x]$ and $F[x] \subseteq \phi(K)[x]$. Hence we can regard f as belonging to both $K[x]$ and $\phi(K)[x]$, and therefore E is a splitting field for f over K and $\phi(K)$. Applying Proposition 11.43 with $F_1 = K$, $F_2 = \phi(K)$ and $E_1 = E_2 = E$ ensures the existence of the desired F -automorphism ψ of E which extends ϕ . $\square$

Corollary 11.75 *Let $E:F$ be a finite normal extension, and suppose that $\alpha_1, \alpha_2 \in E$ are roots of some irreducible polynomial $f \in F[x]$. Then there is an F -automorphism ϕ of F mapping α_1 to α_2 .*

Proof Corollary 11.36 ensures the existence of an F -isomorphism $\psi: F(\alpha_1) \rightarrow F(\alpha_2)$, which can be extended to an F -automorphism of E by Proposition 11.74. $\square$

The extension $\mathbb{Q}(\sqrt[3]{5}):\mathbb{Q}$ in Example 11.65 is not normal, but we can make it into a normal extension by adjoining another element $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$. It so happens that we can do this for any finite extension: we can always adjoin finitely many extra elements to obtain a normal extension containing the original extension field. Not only this, there is effectively a unique choice of minimal normal extension.

Definition 11.76 A **normal closure** of a finite extension $E:F$ is a field K containing E such that $K:F$ is normal, and K contains no proper subfield L such that $L:F$ is normal.

Normal closures always exist, and are unique up to isomorphism:

Proposition 11.77 *Every finite algebraic extension $E:F$ has a normal closure K , which is unique up to F -isomorphism.*

Proof Let $\{\alpha_1, \dots, \alpha_n\}$ be a basis for E considered as a vector space over F , and let $f_i \in F[x]$ be the minimal polynomial for α_i over F , for all $1 \leq i \leq n$. Form the product $f = f_1 \dots f_n \in F[x]$, and let K be its splitting field over F . By Proposition 11.73, the extension $K:F$ is normal. This field K contains all the roots of f and hence all of the roots of its factors $f_1, \dots, f_n$. It therefore also contains the adjoined basis elements $\alpha_1, \dots, \alpha_n$, so K contains E as a subfield.

Now suppose that L is a subfield of K containing E , such that $L:F$ is normal. Since L contains E , it contains all of the adjoined basis elements $\alpha_1, \dots, \alpha_n$. Each α_i is a root of the corresponding minimal polynomial f_i , and since $L:F$ is normal, L must contain all of the roots of f_i , and hence $L = K$. Therefore K is a normal closure of $E:F$.

To prove uniqueness, suppose that K_1 and K_2 are both normal closures of $E:F$. The polynomial $f = f_1 \dots f_n$ splits in both K_1 and K_2 , so each contains a splitting field for f over F . But these splitting fields are each normal closures for $E:F$, and must therefore be equal to K_1 and K_2 respectively. By Corollary 11.44, K_1 and K_2 are F -isomorphic. $\square$

The next proposition will be useful a little later when we study the Galois Correspondence, and prove the Fundamental Theorem of Galois Theory.¹⁷

Proposition 11.78 *Let K be an intermediate subfield of a finite normal extension $E:F$. Then $K:F$ is normal if and only if every F -monomorphism $\phi: K \rightarrow E$ is an F -automorphism of K .*

Proof Suppose that $K:F$ is normal. Let $\phi: K \rightarrow E$ be an F -monomorphism, and let $\alpha \in K$. Let f be the minimal polynomial of α over F , so $f(\alpha) = 0$ and hence $\phi(f(\alpha)) = \phi(0) = 0$. But $\phi(f(\alpha)) = f(\phi(\alpha))$ since ϕ is an F -monomorphism, and in particular $f(\phi(\alpha)) = \phi(f(\alpha)) = 0$, so $\phi(\alpha)$ is also a root of f . Since $K:F$ is normal, $\phi(\alpha)$ must also lie in K . Therefore $\phi(E) \subseteq E$. But $K:F$ is a finite extension, and hence K may be regarded as a finite-dimensional vector space over F . Since ϕ is an F -monomorphism, it follows that $\dim_F \phi(K) = \dim_F K$, so $\phi(K) = K$, and thus ϕ (or, strictly speaking, its restriction $\phi|_K$) is an F -automorphism of K .

Now suppose that every F -monomorphism $\phi: K \rightarrow E$ is an F -auto-

¹⁷ Theorem 11.98, page 463.

morphism of K . Let $f \in F[x]$ be an irreducible polynomial with a root $\alpha \in K$. We want to show that all of the roots of f lie in K . We know that $E:F$ is normal, so let $\beta \in E$ be some other root of f . Then by Corollary 11.75 there is an F -automorphism $\psi: E \rightarrow E$ mapping α to β . The restriction $\psi|_K: K \rightarrow E$ is an F -monomorphism, and by hypothesis is therefore an F -automorphism of K , so $\beta = \psi|_K(\alpha)$ also lies in K . Hence $K:F$ is normal. $\square$

The other situation in which the Galois Correspondence can fail is shown in Example 11.66. There, the problem was that the polynomial $x^p - \alpha$ has only a single repeated root in $\mathbb{F}_p(\alpha^{1/p})$. In order to rule out this situation, we introduce the following definitions.

Definition 11.79 An irreducible polynomial $f \in F[x]$ is **separable** over F if it has no repeated roots in its splitting field. More generally, any polynomial $g \in F[x]$ is **separable** over F if all of its irreducible factors are.

An algebraic element α in some extension $E:F$ is **separable** over F if its minimal polynomial is separable over F .

An algebraic extension $E:F$ is **separable** if every element $\alpha \in E$ is separable over F .

A polynomial, element or extension is said to be **inseparable** if it isn't separable.

This highlights one of the recurring themes in this chapter: the close interrelationship between polynomials, algebraic elements and field extensions. We'll now look at some examples illustrating the concept of separability in each of these contexts.

Example 11.80 The polynomial $x^2+1 \in \mathbb{Q}[x]$ is separable over $\mathbb{Q}$, since it has no repeated roots in its splitting field $\mathbb{Q}(i)$: it factorises as $(x+i)(x-i)$.

Example 11.81 The element $\sqrt{2} \in \mathbb{Q}(\sqrt{2})$ is separable over $\mathbb{Q}$, since its minimal polynomial $x^2-2 \in \mathbb{Q}[x]$ has no repeated roots in its splitting field $\mathbb{Q}(\sqrt{2})$: it factorises as $(x+\sqrt{2})(x-\sqrt{2})$.

Example 11.82 The extension $\mathbb{C}:\mathbb{R}$ is separable. Suppose it isn't, and there exists some $\alpha \in \mathbb{C} \setminus \mathbb{R}$ that is inseparable over $\mathbb{R}$. Then its minimal polynomial $f \in \mathbb{R}[x]$ has repeated roots in $\mathbb{C}$. Since $[\mathbb{C}:\mathbb{R}] = 2$, it follows that $\deg(f) = 1$ or 2 . If $\deg(f) = 1$ then f can have no repeated roots, so $\deg(f)$ must equal 2 . The element α must therefore have multiplicity 2 as a root of f . Furthermore, f can have no other roots, and must therefore factorise as

$$f = (x - \alpha)^2 = x^2 - 2\alpha x + \alpha^2.$$

But $f \in \mathbb{R}[x]$, so all its coefficients must be real. In particular, $-2\alpha = a$ for some $a \in \mathbb{R}$. Hence $\alpha = -a/2 \in \mathbb{R}$. So f is reducible in $\mathbb{R}[x]$, which contradicts our hypothesis that $\alpha \in \mathbb{C} \setminus \mathbb{R}$. Therefore $\mathbb{C} : \mathbb{R}$ is separable.

This last example shows that every algebraic element in $\mathbb{C}$ is separable over $\mathbb{R}$. It also so happens that every polynomial in $\mathbb{R}[x]$ is separable over $\mathbb{R}$. Fields with this property are of particular interest:

Definition 11.83 A field F is **perfect** if every polynomial $f \in F[x]$ is separable over F .

To characterise separable polynomials and perfect fields we need the following proposition.

Proposition 11.84 An irreducible polynomial $f \in F[x]$ is separable over F if either

- (i) $\text{char}(F) = 0$, or
- (ii) $\text{char}(F) = p > 0$ and $f \neq a_0 + a_1x^p + a_2x^{2p} + \cdots + a_nx^{np}$.

Proof If f is inseparable then it has a repeated root in its splitting field. By Proposition 11.50, f and its formal derivative df have some non-constant common factor $g \in F[x]$, so that $g|f$ and $g|df$. But f is irreducible in $F[x]$, so f and g are associates in the sense of Definition 10.15, and hence $f|df$. However, $\deg(df) < \deg(f)$, so $f|df$ only when $df = 0$.

This can't happen if $\text{char}(F) = 0$, since we've tacitly assumed that f isn't a constant (if it is, then it has no roots and questions of separability become irrelevant). If $\text{char}(F) = p > 0$ then the only situation in which a nonconstant polynomial f can have $df = 0$ is if all of its nonzero terms have exponents that are multiples of p . That is, $f = a_0 + a_1x^p + \cdots + a_nx^{np}$, in which case $df = pa_1^{p-1}x^{p-1} + \cdots + npa_nx^{np-1} = 0$. $\square$

Corollary 11.85 If $\text{char}(F) = 0$ then F is perfect.

Proof If $\text{char}(F) = 0$ then any irreducible polynomial $f \in F[x]$ is separable; furthermore any other polynomial is a product of irreducible polynomials and is hence also separable. $\square$

Corollary 11.86 Let $E : F$ be an algebraic extension, and suppose that $\text{char}(F) = 0$. Then $E : F$ is a separable extension.

Proof The minimal polynomial $f_\alpha \in F[x]$ of any element $\alpha \in E$ is separable by Proposition 11.84, so every $\alpha \in E$ is separable, and therefore $E : F$ is a separable extension. $\square$

So, fields of characteristic 0 are perfect: these include $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. But what about fields of characteristic $p > 0$? To answer this question we need to introduce a special field homomorphism named after the

German mathematician Georg Frobenius (1849–1917).

Definition 11.87 Let F be a field with $\text{char}(F) = p > 0$. The **Frobenius map** or **Frobenius monomorphism** is the function $\Phi: F \rightarrow F$ given by $\Phi(a) = a^p$.

Proposition 11.88 *The Frobenius map $\Phi: F \rightarrow F$ is a field monomorphism. If F is finite, then Φ is an automorphism.*

Proof Given any $a, b \in F$, we see that $\Phi(ab) = (ab)^p = a^p b^p = \Phi(a)\Phi(b)$. Also, by the Freshman's Binomial Theorem¹⁸ we have $\Phi(a + b) = (a + b)^p = a^p + b^p = \Phi(a) + \Phi(b)$. Therefore Φ is a field homomorphism.

¹⁸ Proposition 8.20, page 326.

Now suppose that $\Phi(a) = \Phi(b)$. Then $\Phi(a) - \Phi(b) = \Phi(a - b) = 0$. If $a \neq b$ then

$$\Phi(1) = \Phi((a-b)(a-b)^{-1}) = \Phi(a-b)\Phi((a-b)^{-1}) = 0.$$

This is a contradiction, since $\Phi(1) = 1$, so it must instead be the case that $a = b$, and therefore Φ is injective.

If F is finite, then surjectivity follows by a simple counting argument. Since Φ is injective, $|\Phi(F)| = |F|$, so $\Phi(F) = F$. Therefore Φ is surjective and hence an automorphism. $\square$

The Frobenius automorphism enables us to settle the question of when fields of prime characteristic are perfect.

Proposition 11.89 *If $\text{char}(F) = p > 0$ then F is perfect if and only if $\Phi(F) = F$.*

Proof Suppose that $\Phi(F) = F$. If $f \in F[x]$ is irreducible and inseparable over F , then by Proposition 11.84 it must be of the form $a_0 + a_1 x^p + \cdots + a_n x^{np}$. If $F = \Phi(F)$ then for every coefficient a_i there is an element $b_i \in F$ such that $b_i^p = a_i$. Hence

$$\begin{aligned} f &= b_0^p + b_1^p x^p + \cdots + b_n^p x^{np} \\ &= (b_0 + b_1 x + \cdots + b_n x^n)^p \end{aligned}$$

which is reducible. So if $\Phi(F) = F$ then $F[x]$ contains no inseparable polynomials, and hence F is perfect.

Conversely, suppose that $\Phi(F) \neq F$. Then choose some element $a \in F \setminus \Phi(F)$, and consider the polynomial $f = x^p - a \in F[x]$. Suppose that $\alpha^p = a$. Then $x^p - a = x^p - \alpha^p = (x - \alpha)^p$ by the Freshman's Binomial Theorem again, and so f has only a single repeated root in any splitting field over F . All that remains is to show that this polynomial is irreducible over F . If not, then any proper, nontrivial monic factor of f must have the form $(x - \alpha)^m$ for some m such that $0 < m < p$. The coefficient of x^{m-1} in $(x - \alpha)^m$ is $-m\alpha$, so for $x^p - a$

to be reducible over F we require $-m\alpha \in F$. But $m \in \mathbb{F}_p \subseteq F$, so this means that $\alpha \in F$ too, and hence $a = \alpha^p \in \Phi(F)$, which is a contradiction. So if $\Phi(F) \neq F$, there exists at least one irreducible polynomial $x^p - a$ that is inseparable over F , and therefore F isn't perfect. $\square$

Corollary 11.90 *Finite fields are perfect.*

Proof Suppose that F is a finite field. Then $\text{char}(F) = p > 0$ by Proposition 11.51, and $\Phi(F) = F$ by Proposition 11.88. Hence by Proposition 11.89 F is perfect. $\square$

Putting all this together we get the following characterisation of perfect fields:

Proposition 11.91 *A field F is perfect if and only if either*

- (i) $\text{char}(F) = 0$, or
- (ii) $\text{char}(F) = p > 0$ and $\Phi(F) = F$.

As a consequence of this, it is relatively difficult to find good examples of inseparable polynomials and imperfect fields.

These two properties of normality and separability are clearly important, and so we now consider extensions that satisfy them both:

Definition 11.92 A finite field extension $E : F$ is said to be a **Galois extension** if it is both normal and separable.

Proposition 11.93 *Let K be an intermediate subfield of a finite extension $E : F$. Then $E : K$ is normal if $E : F$ is normal, and separable if $E : F$ is separable.*

Proof Suppose that $E : F$ is normal. Then by Proposition 11.73 there is a polynomial $f \in F[x]$ with splitting field E over F . Since $F \subseteq K$ it follows that $F[x] \subseteq K[x]$, and hence E is also a splitting field for f over K , and so $E : K$ is normal.

Now suppose that $E : F$ is separable, and consider some element $\alpha \in E$ with minimal polynomial f_F over F and f_K over K . If f_F is separable, then $f_K | f_F$ in $K[x]$. Suppose that f_K isn't separable. Then by Proposition 11.50 there is a non-constant polynomial $g \in K[x]$ that divides both f_K and df_K . Since $f_K | f_F$ it follows that g divides both f_F and df_F . But this can't happen, because f_F is separable. Hence f_K must also be separable, and therefore the extension $E : K$ must be separable. $\square$

In several of our key examples it has been the case that the degree $[E : F]$ of an extension is equal to the order $|\text{Gal}(E : F)|$ of its Galois group. This doesn't always happen, as we've seen in our key counterexamples, but it does if we insist on the now obvious nice properties of normality and separability.

Proposition 11.94 *Let N be a normal closure of a finite separable extension $E:F$. Then there are exactly $[E:F]$ distinct F -monomorphisms from E into N .*

Proof We prove this by induction on the degree $[E:F]$ of the extension. If $[E:F] = 1$ then $E = F = N$ and the only F -monomorphism $\phi: E \rightarrow N$ is the identity map id_F .

Now suppose that $k > 1$ and the proposition holds for all extensions $E:F$ with degree $[E:F] < k$. Choose some arbitrary element $\alpha \in E \setminus F$ and let f be its minimal polynomial over F . Then by Proposition 11.30 we have $[F(\alpha):F] = \deg(f) = r > 1$. This polynomial f is irreducible, and has one root $\alpha \in E \subseteq N$, so since $N:F$ is normal, f splits completely in N . Its zeros $\alpha_1 = \alpha, \dots, \alpha_r$ all lie in N , and by separability are distinct. By the induction hypothesis, there are exactly $s = k/r = [E:F(\alpha)]$ distinct $F(\alpha)$ -monomorphisms $\phi_1, \dots, \phi_s: E \rightarrow N$.

By Corollary 11.75 there are r distinct F -automorphisms $\psi_1, \dots, \psi_r$ of N with $\psi_i(\alpha_1) = \alpha_i$ for $1 \leq i \leq r$. We now compose these with the maps $\phi_1, \dots, \phi_s$ to obtain rs F -monomorphisms

$$\chi_{i,j}: E \rightarrow N; \quad \chi_{i,j}(a) = \psi_i(\phi_j(a))$$

for all $a \in E$, and such that $1 \leq i \leq r$ and $1 \leq j \leq s$.

We now need to show that these F -monomorphisms are all distinct, and then that any F -monomorphism from E to N is of this form.

Consider two such F -monomorphisms $\chi_{m,n}$ and $\chi_{p,q}$ for some values of m, n, p and q . For these to be equal, we require that $\chi_{m,n}(a) = \chi_{p,q}(a)$ for all $a \in E$. In particular,

$$\begin{aligned} \chi_{m,n}(\alpha) &= \psi_m(\phi_n(\alpha)) = \psi_m(\alpha) = \alpha_m \\ \text{and} \quad \chi_{p,q}(\alpha) &= \psi_p(\phi_q(\alpha)) = \psi_p(\alpha) = \alpha_p. \end{aligned}$$

These are clearly equal only when $m = p$. Furthermore, if $\chi_{m,n}(a) = \chi_{m,q}(a)$, then $\psi_m(\phi_n(a)) = \psi_m(\phi_q(a))$. Since ψ_m is injective, it follows that $\phi_n(a) = \phi_q(a)$. Hence $\chi_{m,n} = \chi_{m,q}$ only when $\phi_n(a) = \phi_q(a)$ for all $a \in E$. But the $F(\alpha)$ -monomorphisms $\phi_1, \dots, \phi_r$ are all distinct, so this can only happen if $n = q$. The F -automorphisms $\chi_{i,j}$ are therefore all distinct.

Now let χ be any F -automorphism from E to N . It must map the root α to some root α_i of f in N , so define $\phi(a) = \psi_i^{-1}(\chi(a))$ for all $a \in E$. In particular, $\phi(\alpha) = \psi_i^{-1}(\chi(\alpha)) = \psi_i^{-1}(\alpha_i) = \alpha$, so ϕ is an $F(\alpha)$ -monomorphism. Earlier we deduced from the induction hypothesis that there are exactly $s = k/r$ $F(\alpha)$ -monomorphisms $\phi_1, \dots, \phi_s: E \rightarrow N$, and so this new map ϕ must be one of them, say ϕ_j . Hence $\phi_j = \psi_i^{-1} \circ \chi$, so $\chi = \psi_i \circ \phi_j = \chi_{i,j}$. Therefore every F -automorphism is

of the required form. This completes the induction step, and thereby the proof. $\square$

Corollary 11.95 *If $E:F$ is a Galois extension then $|\text{Gal}(E:F)| = [E:F]$.*

Proof If $E:F$ is a Galois extension, then it is finite, normal and separable, and its normal closure is E itself. By the above proposition, there are precisely $[E:F]$ F -monomorphisms from $E \rightarrow E$. By Proposition 11.78 these are all actually F -automorphisms. The Galois group $\text{Gal}(E:F)$ consists of all of these F -automorphisms, and hence $|\text{Gal}(E:F)| = [E:F]$ as claimed. $\square$

Proposition 11.96 *If $E:F$ is a Galois extension, then $\mathcal{F}(\text{Gal}(E:F))=F$.*

Proof Observe that $\text{Gal}(E:F) = \mathcal{G}(F)$. By Proposition 11.67 we have $F \subseteq \mathcal{F}(\mathcal{G}(F))$. It follows from Corollary 11.95 that

$$|\mathcal{G}(F)| = |\text{Gal}(E:F)| = [E:F].$$

Furthermore, by Proposition 11.68

$$[E:\mathcal{F}(\mathcal{G}(F))] = |\text{Gal}(E:F)| = [E:F].$$

By the Tower Law,¹⁹

$$[E:F] = [E:\mathcal{F}(\mathcal{G}(F))][\mathcal{F}(\mathcal{G}(F)):F],$$

so $[\mathcal{F}(\mathcal{G}(F)):F] = 1$, and thus $\mathcal{F}(\mathcal{G}(F)) = F$. $\square$

The converse of this is also true, but the proof is somewhat involved, and we won't need it for the main aim of this section.

Proposition 11.97 *Let K be an intermediate subfield of a Galois extension $E:F$. If $\phi \in \text{Gal}(E:F)$ then $\mathcal{G}(\phi(K)) = \phi\mathcal{G}(K)\phi^{-1}$.*

Proof Let $\psi \in \mathcal{G}(\phi(K)) = \text{Gal}(E:\phi(K))$ and let $a \in K$ be any element of the intermediate field. Then $\phi(a) \in \phi(K)$, and ψ fixes $\phi(a)$; that is, $\psi(\phi(a)) = \phi(a)$. Therefore

$$(\phi^{-1} \circ \psi \circ \phi)(a) = \phi^{-1}(\psi(\phi(a))) = \phi^{-1}(\phi(a)) = a$$

and hence $\phi^{-1} \circ \psi \circ \phi \in \text{Gal}(E:K) = \mathcal{G}(K)$. So $\phi^{-1}\mathcal{G}(\phi(K))\phi \leq \mathcal{G}(K)$, which implies that $\phi\mathcal{G}(K)\phi^{-1} \leq \mathcal{G}(\phi(K))$.

Now let $\psi \in \mathcal{G}(K) = \text{Gal}(E:K)$, let $b \in \phi(K)$, and let $a = \phi^{-1}(b) \in K$. Since ψ fixes all the elements of K , in particular we have $\psi(a) = a$. Then

$$(\phi \circ \psi \circ \phi^{-1})(b) = (\phi \circ \psi \circ \phi^{-1})(\phi(a)) = \phi(\psi(a)) = \phi(a) = b,$$

so $\phi \circ \psi \circ \phi^{-1}$ fixes any element of $\phi(K)$, and is hence an element of $\text{Gal}(E:\phi(K)) = \mathcal{G}(\phi(K))$. Thus $\phi\mathcal{G}(K)\phi^{-1} \leq \mathcal{G}(\phi(K))$, and therefore $\phi\mathcal{G}(K)\phi^{-1} = \mathcal{G}(\phi(K))$ as claimed. $\square$

¹⁹ Proposition 11.21, page 428.

We are now ready to state and prove one of the main results of Galois theory. This theorem answers the question we asked earlier about when the maps $\mathcal{F}$ and $\mathcal{G}$ are mutually inverse, and also tells us some other useful facts about Galois groups.

Theorem 11.98 (Fundamental Theorem of Galois Theory) *Let $E:F$ be a Galois extension, and suppose that K is an intermediate subfield of E , and H is a subgroup of the Galois group $\text{Gal}(E:F)$. Then:*

- (i) $\mathcal{F}(\mathcal{G}(K)) = K$ and $\mathcal{G}(\mathcal{F}(H)) = H$;
- (ii) $|\mathcal{G}(K)| = [E:K]$ and $|\text{Gal}(E:F)|/|\mathcal{G}(K)| = [K:F]$;
- (iii) K is a normal extension of F if and only if $\mathcal{G}(K) \trianglelefteq \text{Gal}(E:F)$; and
- (iv) if K is a normal extension of F then $\text{Gal}(K:F) \cong \text{Gal}(E:F)/\mathcal{G}(K)$.

Proof By Proposition 11.93, $E:K$ is a Galois extension, and hence

$$\mathcal{F}(\mathcal{G}(K)) = \mathcal{F}(\text{Gal}(E:K)) = K \quad (11.18)$$

by Proposition 11.96. And

$$|\mathcal{G}(K)| = |\text{Gal}(E:K)| = [E:K]$$

by Corollary 11.95.

Setting $K = \mathcal{F}(H)$ in (11.18), we get $\mathcal{F}(\mathcal{G}(\mathcal{F}(H))) = \mathcal{F}(H)$. Proposition 11.68 says that $|H| = [E:\mathcal{F}(H)]$, and hence

$$|H| = [E:\mathcal{F}(H)] = [E:\mathcal{F}(\mathcal{G}(\mathcal{F}(H)))] = |\mathcal{G}(\mathcal{F}(H))|.$$

By Proposition 11.67 we have $H \leq \mathcal{G}(\mathcal{F}(H))$, and since both H and $\mathcal{G}(\mathcal{F}(H))$ are finite groups, it must follow that $\mathcal{G}(\mathcal{F}(H)) = H$ as claimed.

By the Tower Law,²⁰

²⁰ Proposition 11.21, page 428.

$$[E:F] = [E:K][K:F]$$

and hence

$$[K:F] = [E:F]/[E:K].$$

Applying Corollary 11.95, we have $[E:F] = |\text{Gal}(E:F)|$ and $[E:K] = |\text{Gal}(E:K)| = |\mathcal{G}(K)|$, since both $E:F$ and $E:K$ are Galois extensions. Thus

$$[K:F] = |\text{Gal}(E:F)|/|\mathcal{G}(K)|.$$

This completes the first two parts of the proof.

Now suppose that $K:F$ is a normal extension, and let $\phi \in \text{Gal}(E:F)$ be some arbitrary F -automorphism of E . Then the restriction $\phi|_K$ to K is an F -monomorphism from K to E , and hence by Proposition 11.78 is actually an F -automorphism of K , so $\phi(K) = K$. Then from Proposition 11.97 we have

$$\phi\mathcal{G}(K)\phi^{-1} = \mathcal{G}(\phi(K)) = \mathcal{G}(K),$$

so $\mathcal{G}(K)$ is closed under conjugation by any element of $\text{Gal}(E:F)$, and therefore $\mathcal{G}(K) \trianglelefteq \text{Gal}(E:F)$.

Conversely, suppose that $\mathcal{G}(K) \trianglelefteq \text{Gal}(E:F)$. Let $\psi: K \rightarrow E$ be any F -monomorphism from K into E . By Proposition 11.74 this can be extended to an F -automorphism ϕ of E , such that $\phi|_K = \psi$. Since $\mathcal{G}(K)$ is normal in $\text{Gal}(E:F)$ it is closed under conjugation by ϕ , and so $\phi\mathcal{G}(K)\phi^{-1} = \mathcal{G}(K)$. From Proposition 11.97 we deduce that $\mathcal{G}(\phi(K)) = \mathcal{G}(K)$, and by (11.18) it follows that

$$\phi(K) = \mathcal{F}(\mathcal{G}(\phi(K))) = \mathcal{F}(\mathcal{G}(K)) = K.$$

Therefore $\psi(K) = \phi|_K(K) = K$, so ψ is an F -automorphism of K , and by Proposition 11.78 $K:F$ is normal. This completes part (iii).

To prove part (iv), suppose that $K:F$ is a normal extension. Let $f: \text{Gal}(E:F) \rightarrow \text{Gal}(K:F)$ be defined by $f(\phi) = \phi|_K$ for any F -automorphism ϕ of E . We must first confirm that this is a well-defined group homomorphism. Let $\phi \in \text{Gal}(E:F)$ be any F -automorphism of E . Then $\phi|_K$ is an F -monomorphism, and by Proposition 11.78 it is an F -automorphism of K , and thus lies in $\text{Gal}(K:F)$. Furthermore, for any $\phi, \psi \in \text{Gal}(E:F)$ we have

$$f(\phi \circ \psi) = (\phi \circ \psi)|_K = \phi|_K \circ \psi|_K = f(\phi) \circ f(\psi),$$

so f is a group homomorphism. By Proposition 11.74, since $K:F$ is normal, any F -automorphism of K is the restriction of some F -automorphism of E . The homomorphism f is thus surjective.

The kernel $\ker(f)$ consists of all F -automorphisms of E that restrict to the identity on K . But this is precisely $\mathcal{G}(K) = \text{Gal}(E:K)$. Applying the First Isomorphism Theorem for groups,²¹ we obtain

$$\text{Gal}(K:F) = \text{im}(f) \cong \text{Gal}(E:F) / \ker(f) = \text{Gal}(E:F) / \mathcal{G}(K)$$

as claimed. $\square$

²¹ Theorem 4.40, page 117.

Radical, n. A miscreant who would forestall the future by discrediting the past and abolishing the present.

— Ambrose Bierce (1842–c.1914),
The New York American, 29 June 1906,
 in: *The Unabridged Devil's Dictionary*
 (2000)

²² J Friberg, *A Geometric Algorithm with Solutions to Quadratic Equations in a Sumerian Juridical Document from Ur III Umma*, Cuneiform Digital Library Journal 2009.3 (2009).

11.6 Solving equations by radicals

METHODS FOR SOLVING QUADRATIC EQUATIONS have been known for thousands of years: the earliest known examples occur in Sumerian cuneiform tablets dating from about 2000 BC.²² The standard formula that most of us learn at school can be derived by a method called **completing the square**. Given

$$ax^2 + bx + c = 0$$

where $a, b, c \in \mathbb{Q}$ and $a \neq 0$, we first divide through by a to get a monic equation

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0.$$

The trick is then to rewrite as much of the left hand side as possible in the form $(x + d)^2$ for some appropriate $d \in \mathbb{Q}$. Setting $d = \frac{b}{2a}$, we get

$$(x + d)^2 = (x + \frac{b}{2a})^2 = x^2 + \frac{b}{a}x + \frac{b^2}{4a^2},$$

which is very nearly what we want: it's just the constant term that needs addressing. So

$$x^2 + \frac{b}{a}x + \frac{c}{a} = (x + \frac{b}{2a})^2 - \frac{b^2}{4a^2} + \frac{c}{a} = 0,$$

which we can then rearrange to get

$$(x + \frac{b}{2a})^2 = \frac{b^2 - 4ac}{4a^2}.$$

Taking square roots and solving for x we get the familiar quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Explicit formulae for solving cubic and quartic equations were developed in Italy during the 16th century, amid an environment of fiercely counterproductive competition. These formulae are rather cumbersome, but they all only require basic arithmetic operations (addition, subtraction, multiplication and division) together with finite roots.

This isn't true for all polynomial equations, and to understand why will involve the algebraic machinery developed in this chapter together with some of the advanced group theory we studied in Chapter 7.

We start by identifying a special class of algebraic field extensions.

Definition 11.99 A field extension $E:F$ is a **radical extension** if there exists a sequence

$$F = K_0, K_1, \dots, K_m = E$$

of fields such that $K_{i+1} = K_i(\alpha_i)$ where α_i is a root of a polynomial f_i of the form $x^{n_i} - a_i$ in $K_i[x]$, with $0 \leq i < m$ and $n_i \in \mathbb{N}$.

What this means in practice is that radical extensions are exactly those of the form $F(\alpha):F$ where α can be constructed from elements of F using just the four basic arithmetical operations and finite roots.

Example 11.100 Let $\alpha = (1 + \sqrt{3})^{1/5} - \sqrt[3]{5}(\frac{1}{2} + \sqrt{2})^{1/7}$. Then $\mathbb{Q}(\alpha):\mathbb{Q}$ is a radical extension, since

$$\begin{aligned} K_0 &= \mathbb{Q}; & K_1 &= K_0(\alpha_0), & f_0 &= \alpha_0^2 - 3; \\ K_2 &= K_1(\alpha_1), & f_1 &= \alpha_1^5 - (1 + \alpha_0); & K_3 &= K_2(\alpha_2), & f_2 &= \alpha_2^3 - 5; \\ K_4 &= K_3(\alpha_3), & f_3 &= \alpha_3^2 - 2; & K_5 &= K_4(\alpha_4), & f_4 &= \alpha_4^7 - (2 + \alpha_3). \end{aligned}$$

By the Tower Law, $[\mathbb{Q}(\alpha):\mathbb{Q}] = \prod_{i=0}^4 n_i = 2 \times 5 \times 3 \times 2 \times 7 = 420$.

Our aim is to identify a privileged class of polynomials whose roots lie in radical extension fields. The key is to consider splitting fields:

Definition 11.101 A polynomial $f \in F[x]$ is **soluble by radicals** if there exists a radical extension E of F which contains, or is itself, a splitting field for f .

A given polynomial of degree five or greater is not necessarily soluble by radicals. The proof is originally due to the Italian mathematician and physician Paolo Ruffini (1765–1822) and the Norwegian mathematician Niels Henrik Abel (1802–1829). The more general theory was developed by the French mathematician Évariste Galois (1811–1832).

Recall from Proposition 11.46 that a field must have either zero or prime characteristic. We will only consider the zero-characteristic case here. The essential idea is that the Galois group encodes fundamental information about the field extension or polynomial under investigation. In particular, a given polynomial is soluble by radicals exactly when its Galois group is soluble in the sense of Definition 7.36. We'll prove this in two halves.

Proposition 11.102 Let $f \in F[x]$ be a polynomial over a field F with characteristic zero. If the Galois group $\text{Gal}(f, F)$ is soluble, then f is soluble by radicals.

In order to prove this, we need a few preliminary results. First we briefly study extensions with cyclic Galois groups.

Definition 11.103 Let F be a field of characteristic zero. A Galois extension $E:F$ is **cyclic** if its Galois group is cyclic; that is, if $\text{Gal}(E:F) \cong \mathbb{Z}_n$ for some $n \in \mathbb{N}$.

Definition 11.104 Let $E:F$ be a finite extension of a field F of characteristic zero. Let N be a normal closure of E . By Proposition 11.94 there are $[E:F]$ distinct F -monomorphisms from E into N . We define the **norm** $N_{E:F}$ and **trace** $\text{Tr}_{E:F}$ of $E:F$ to be

$$N_{E:F}(a) = \prod_{\phi \in \text{Gal}(E:F)} \phi(a) \quad \text{and} \quad \text{Tr}_{E:F}(a) = \sum_{\phi \in \text{Gal}(E:F)} \phi(a)$$

for all $a \in E$.

The first important result we will prove is due to David Hilbert (1862–1943) and Ernst Kummer (1810–1893).²³

Proposition 11.105 (Hilbert's Theorem 90) Let $E:F$ be a cyclic extension, and let τ be a generator of the Galois group $\text{Gal}(E:F)$. Then for any element $a \in E$, the norm $N_{E:F}(a) = 1$ if and only if there is an element $b \in E$ such that $a = b/\tau(b)$.

Proof Suppose that $N_{E:F}(a) = a\tau(a)\tau^2(a) \dots \tau^{n-1}(a) = 1$, so

$$a^{-1} = \tau(a)\tau^2(a) \dots \tau^{n-1}(a). \quad (11.19)$$

²³ This result is often referred to as 'Hilbert's Theorem 90', because it is the ninetieth theorem stated in *Zahlbericht*, his 1897 report on algebraic number theory; it had, however, been proved in 1855 by Ernst Kummer (1810–1893).

Let

$$d_k = \prod_{i=0}^k \tau^i(a) = a\tau(a)\tau^2(a) \dots \tau^{n-1}(a)$$

for $0 \leq k < n$. In particular, $d_{n-1} = N_{E:F}(a) = 1$ and $d_k = a\tau(d_{k-1})$ for $0 < k < n$.

By Lemma 11.69 the automorphisms $\text{id}, \tau, \tau^2, \dots, \tau^{n-1}$ are linearly independent over E , so

$$\phi = d_0 \text{id} + d_1 \tau + d_2 \tau^2 + \dots + d_{n-1} \tau^{n-1}$$

isn't the zero map. In particular, there exists some $c \in E$ such that

$$\phi(c) = d_0 c + d_1 \tau(c) + \dots + d_{n-1} \tau^{n-1}(c) \neq 0.$$

Let $b = \phi(c)$, and apply τ to get

$$\begin{aligned} \tau(b) &= \tau(d_0 c + \dots + d_{n-1} \tau^{n-1}(c)) \\ &= \tau(d_0) \tau(c) + \tau(d_1) \tau^2(c) + \dots + \tau(d_{n-1}) \tau^n(c) \\ &= \tau(d_{n-1}) c + \tau(d_0) \tau(c) + \dots + \tau(d_{n-2}) \tau^{n-1}(c) \\ &= (d_0 c + d_1 \tau(c) + \dots + d_{n-1} \tau^{n-1}(c)) / a \\ &= b / a, \end{aligned}$$

and hence $a = b / \tau(b)$.

Conversely, suppose there exists some $b \in E$ with $a = b / \tau(b)$. Then

$$N_{E:F}(a) = a\tau(a)\tau^2(a) \dots \tau^{n-1}(a) = \frac{b}{\tau(b)} \frac{\tau(b)}{\tau^2(b)} \dots \frac{\tau^{n-1}(b)}{\tau^n(b)} = 1$$

as claimed. $\square$

Example 11.106 The extension $\mathbb{C}:\mathbb{R}$ is cyclic, with Galois group $\text{Gal}(\mathbb{C}:\mathbb{R}) \cong \mathbb{Z}_2$ generated by the conjugation automorphism $\tau: z \mapsto \bar{z}$. The norm $N_{\mathbb{C}:\mathbb{R}}$ is defined by $N_{\mathbb{C}:\mathbb{R}}(z) = z\tau(z) = z\bar{z} = |z|^2$ for any $z \in \mathbb{C}$. The complex numbers of unit norm are therefore exactly those with modulus 1.

Suppose $z = \langle 1, \theta \rangle$ in polar form, with $\theta \in (-\pi, \pi]$. We can then choose $w = \langle r, \theta/2 \rangle$ for any $r > 0$ and get

$$w / \tau(w) = w / \bar{w} = \langle r, \theta/2 \rangle / \langle r, -\theta/2 \rangle = \langle 1, \theta \rangle.$$

Conversely, any nonzero complex number w yields a complex number $z = w / \tau(w)$ of unit modulus.

Proposition 11.107 Suppose that $f = x^m - 1 \in F[x]$ splits completely over F , a field with characteristic zero. Let $E:F$ be a cyclic extension with $[E:F] = m$. Then there exists an element $a \in F$ for which $g = x^m - a$ is irreducible over F , and E is a splitting field for g over F generated by a single root of g .

Proof Let τ be a generator of $\text{Gal}(E:F) \cong \mathbb{Z}_m$. Since $f = x^m - 1$ splits completely over F , there exists a primitive m th root of unity $\omega \in F$, which is fixed by every F -automorphism of E . That is, $\tau^k(\omega) = \omega$ for all $k = 0, \dots, m-1$, and hence

$$N_{E:F}(\omega) = \prod_{k=0}^{m-1} \tau^k(\omega) = \prod_{k=0}^{m-1} \omega = \omega^m = 1.$$

²⁴ Proposition 11.105, page 466.

By Hilbert's Theorem 90,²⁴ there exists some $\alpha \in E$ with $\omega = \alpha/\tau(\alpha)$, and hence $\tau(\alpha) = \omega^{-1}\alpha$. This means that $\alpha \in E \setminus F$, and also that

$$\tau^k(\alpha) = \omega^{-k}\alpha \quad \text{for } k = 0, \dots, m-1.$$

This means that the only F -automorphism that fixes $F(\alpha)$ is the identity, so $\mathcal{G}(F(\alpha)) = \{\text{id}\}$.

Since $E:F$ is cyclic, and thus by definition a Galois extension, we can apply the Fundamental Theorem of Galois Theory²⁵ to show that

²⁵ Theorem 11.98, page 463.

$$F(\alpha) = \mathcal{F}(\mathcal{G}(F(\alpha))) = \mathcal{F}(\{\text{id}\}) = E,$$

so E is generated over F by α .

Observe that

$$\tau(\alpha^m) = \tau(\alpha)^m = (\omega^{-1}\alpha)^m = \omega^{-m}\alpha^m = \alpha^m,$$

so α^m is fixed by τ and all its powers, and therefore by every element of $\text{Gal}(E:F)$, so it lies in $\mathcal{F}(\text{Gal}(E:F)) = F$. Let $a = \alpha^m$. Then α is a root of $g = x^m - a$. We now want to show that g is irreducible over F , and that E is its splitting field.

Since α is a root of g , the minimal polynomial for α over F must be a factor of g . But $[E:F] = [F(\alpha):F] = \deg(g) = m$, so the minimal polynomial of α over F must be g itself. By Proposition 11.29, g must be irreducible over F . Finally, since

$$g = x^m - a = (x - \alpha)(x - \omega\alpha) \dots (x - \omega^{m-1}\alpha),$$

which splits completely over $E = F(\alpha)$ and not over any subfield of E , it follows that E is a splitting field for g over F . $\square$

The next proposition is a technical result about extensions of splitting fields.

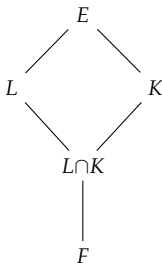


Figure 11.4: Subfield lattice for Proposition 11.108

Proposition 11.108 Let $f \in F[x]$ be a polynomial over a field F of characteristic zero. Let K be a splitting field for f over F , let L be a field containing F , and let E be a splitting field of f over L . Then E contains a subfield isomorphic to K , and $\text{Gal}(E:L) \cong \text{Gal}(K:L \cap K)$.

Proof Figure 11.4 depicts the subfield lattice for this situation.

Since $F \subseteq L$ and $L \subseteq E$ it follows that E is an extension of F . By Definition 11.38, K is the smallest possible field over which $f \in F[x]$ splits completely, so any other field over which f splits completely

must contain a subfield isomorphic to K . Hence we can regard E as an extension of K .

Let $\phi \in \text{Gal}(E:L)$ be some L -automorphism of E . The restriction $\psi = \phi|_K$ is a field monomorphism from K to E . As the restriction of an L -automorphism, ψ must necessarily fix all the elements of L that lie in K , so it is a $(K \cap L)$ -monomorphism. We want to show that this is actually a $(K \cap L)$ -automorphism of K .

By Proposition 11.59, ϕ , as an L -automorphism of E , permutes the roots $\alpha_1, \dots, \alpha_n$ of f in $E = L(\alpha_1, \dots, \alpha_n)$. Its restriction ψ also permutes all of these roots in $K = F(\alpha_1, \dots, \alpha_n)$, while fixing all the elements of F . Hence ψ is an automorphism of K that, as previously noted, fixes not just F but $K \cap L$, so $\psi \in \text{Gal}(K:K \cap L)$.

Restriction to K yields a map $\theta: \text{Gal}(E:L) \rightarrow \text{Gal}(K:K \cap L)$ such that $\theta(\phi) = \phi|_K$, and we now want to show that this is an isomorphism. It is a homomorphism, since for any $\phi_1, \phi_2 \in \text{Gal}(E:L)$ we have

$$\theta(\phi_1 \circ \phi_2) = (\phi_1 \circ \phi_2)|_K = (\phi_1|_K) \circ (\phi_2|_K) = \theta(\phi_1) \circ \theta(\phi_2).$$

It is injective, since if $\theta(\phi_1) = \theta(\phi_2)$ then both $\theta(\phi_1) = \phi_1|_K$ and $\theta(\phi_2) = \phi_2|_K$ permute the roots $\alpha_1, \dots, \alpha_n$ the same, so $\phi_1 = \phi_2$.

We now just need to show that θ is surjective. Suppose that $\phi \in \text{Gal}(E:L)$. Certainly $\theta(\phi)$ fixes $K \cap L$, so $\theta(\phi) \in \text{Gal}(K:K \cap L)$ and $\theta(\text{Gal}(E:L)) \subseteq \text{Gal}(K:K \cap L)$. Now let $a \in K \setminus K \cap L$, so $a \notin L$. There is an L -automorphism $\chi \in \text{Gal}(E:L)$ for which $\chi(a) \neq a$. Hence $\theta(\chi)(a) \neq a$, so a is not fixed by all elements of $\theta(\text{Gal}(E:L))$. Thus $\theta(\text{Gal}(E:L))$ fixes every element of $K \cap L$ and doesn't fix any element of $K \setminus K \cap L$, so $\theta(\text{Gal}(E:L)) = \text{Gal}(K:K \cap L)$ as claimed. $\square$

Proof of Proposition 11.102 Let K be a splitting field for f over F . Suppose $\text{Gal}(f, F) = \text{Gal}(K:F)$ is soluble, and that $|\text{Gal}(K:F)| = m$.

There are two cases to consider, depending on whether F contains an m th root of unity or not.

Case 1 Suppose that F does contain a primitive m th root of unity ω ; it also necessarily contains all m th roots of unity, these being powers of ω . Since $G = \text{Gal}(f, F) = \text{Gal}(K:F)$ is soluble, there exists a series

$$\{\text{id}\} = G_0 \triangleleft G_1 \triangleleft \cdots \triangleleft G_k = G$$

with each quotient $G_i/G_{i-1} \cong \mathbb{Z}_{m_i}$ cyclic, for $0 < i \leq k$.

By the Fundamental Theorem of Galois Theory²⁶ there is a corresponding sequence ²⁶ Theorem 11.98, page 463.

$$F = F_k \subseteq F_{k-1} \subseteq \cdots \subseteq F_1 \subseteq F_0 = E$$

of subfields of E , with $\mathcal{G}(F_i) = \text{Gal}(E:F_i) = G_i$ and $\text{Gal}(F_{i-1}:F_i) = G_i/G_{i-1} \cong \mathbb{Z}_{m_i}$, so $F_{i-1}:F_i$ is a cyclic extension of degree m_i for $0 < i \leq k$.

Furthermore, F_i contains $F_k = E$, which contains all m th roots of unity. Since $m_i | m$, it follows that E , and hence F_i , also contains all m_i th roots of unity, since these are also powers of ω . By Proposition 11.107 there exists some element $\alpha_i \in F_i$ such that $F_i = F_{i+1}(\alpha_i)$ is the splitting field over F_{i+1} of some polynomial $g_i = x^{m_i} - a_i \in F_{i+1}[x]$ where $a_i = \alpha_i^{m_i}$. Therefore f is soluble by radicals.

Case 2 Now suppose that F doesn't contain a primitive m th root of unity. The polynomial $x^m - 1 \in F[x]$ is obviously soluble by radicals. Let ω be a primitive m th root of unity, and adjoin it to F to get $L = F(\omega)$, the splitting field of $x^m - 1$ over F . Now let E be the splitting field of f over L . By Proposition 11.108, E is an extension of K , and $\text{Gal}(E:L) \cong \text{Gal}(K:L \cap K)$, which is a subgroup of $\text{Gal}(K:F)$, which is soluble. Proposition 7.38 says that subgroups of soluble groups are soluble, and hence $\text{Gal}(E:L)$ is soluble. This now reduces to Case 1, and therefore f is again soluble by radicals. $\square$

The converse is also true: polynomials that are soluble by radicals have soluble Galois groups.

Proposition 11.109 *Let $f \in F[x]$ be a polynomial over a field F of characteristic zero. If f is soluble by radicals, then its Galois group $\text{Gal}(f, F)$ is soluble.*

To prove this we first need the following result.

Proposition 11.110 *Let F be a field of characteristic zero, and let $f = x^n - a \in F[x]$. The Galois group $\text{Gal}(f, F)$ is soluble.*

Proof We split the proof into two cases, according to whether F contains an n th root of unity. Let E be the splitting field of f over F .

Case 1 Suppose that F contains a primitive n th root of unity ω . Since F necessarily contains all powers of ω , it thereby contains all n th roots of unity.

Let α be a root of $f = x^n - a$ in E . Then the roots of f are

$$\alpha, \alpha\omega, \alpha\omega^2, \dots, \alpha\omega^{n-1}$$

and $E = F(\alpha)$. Any F -automorphism of E is thus determined entirely by its action on α , and must fix ω .

By Proposition 11.59, any such F -automorphism of E must permute the roots of f in E , in particular sending α to an element of the form $\alpha\omega^i$. Let $\phi, \psi \in \text{Gal}(f, F) = \text{Gal}(E:F)$, with $\phi(\alpha) = \alpha\omega^i$ and $\psi(\alpha) = \alpha\omega^j$. Then

$$(\phi \circ \psi)(\alpha) = \phi(\psi(\alpha)) = \phi(\alpha\omega^j) = \phi(\alpha)\phi(\omega)^j = \alpha\omega^i\omega^j = \alpha\omega^{i+j}$$

$$(\psi \circ \phi)(\alpha) = \psi(\phi(\alpha)) = \psi(\alpha\omega^i) = \psi(\alpha)\psi(\omega)^i = \alpha\omega^j\omega^i = \alpha\omega^{i+j}$$

so $\phi\psi = \psi\phi$, hence $\text{Gal}(E:F)$ is abelian, and therefore soluble by

Proposition 7.42.

Case 2 Suppose instead that F doesn't contain a primitive n th root of unity ω , and adjoin one to get $F(\omega)$, which is the splitting field of $x^n - 1$ over F . Since the roots of f are, as noted in Case 1,

$$\alpha, \alpha\omega, \alpha\omega^2, \dots, \alpha\omega^{n-1},$$

we know that E contains $F(\omega)$.

Let $\phi, \psi \in \text{Gal}(F(\omega):F)$ be two F -automorphisms of $F(\omega)$. Each is completely determined by its action on ω so suppose that $\phi(\omega) = \omega^i$ and $\psi(\omega) = \omega^j$. Then

$$\begin{aligned} (\phi \circ \psi)(\omega) &= \phi(\psi(\omega)) = \phi(\omega^j) = \phi(\omega)^j = (\omega^i)^j = \omega^{ij} \\ (\psi \circ \phi)(\omega) &= \psi(\phi(\omega)) = \psi(\omega^i) = \psi(\omega)^i = (\omega^j)^i = \omega^{ji} \end{aligned}$$

so $\text{Gal}(F(\omega):F)$ is abelian and hence soluble, again by Proposition 7.42.

By the same argument used in Case 1, with $F(\omega)$ replacing F , we see that $\text{Gal}(E:F(\omega))$ is also abelian and therefore soluble.

Since $F(\omega)$ is the splitting field for $x^n - 1$ over F , the extension $F(\omega):F$ is normal by Proposition 11.73. Hence by part (iv) of the Fundamental Theorem of Galois Theory,²⁷

²⁷ Theorem 11.98, page 463.

$$\text{Gal}(F(\omega):F) \cong \text{Gal}(E:F) / \text{Gal}(E:F(\omega)).$$

By Proposition 7.40, since both $\text{Gal}(F(\omega):F)$ and $\text{Gal}(E:F(\omega))$ are both soluble, so is $\text{Gal}(f, F) = \text{Gal}(E:F)$. $\square$

Proof of Proposition 11.109 Let E be the splitting field for f over F , and suppose that $E \subseteq F(\alpha_1, \dots, \alpha_k)$ where $\alpha_1^{n_1} \in F$ and $\alpha_i^{n_i} \in F(\alpha_1, \dots, \alpha_{i-1})$ for some $n_i \in \mathbb{N}$ and $i = 2, \dots, k$.

We proceed by induction on k . For the base case $k = 1$, the splitting field E is contained in $F(\alpha_1)$. Let $a = \alpha_1^{n_1}$, and suppose that K is the splitting field for the polynomial $g = x^{n_1} - a$ over F . Then K contains $F(\alpha_1)$, so $E \subseteq K$.

Both E and K are splitting fields for polynomials over F , and hence by Proposition 11.73 the extensions $E:F$ and $K:F$ are normal, and so by part (iv) of the Fundamental Theorem of Galois Theory²⁸

²⁸ Theorem 11.98, page 463.

$$\text{Gal}(f, F) = \text{Gal}(E:F) \cong \text{Gal}(K:F) / \text{Gal}(K:E).$$

By Proposition 11.110, $\text{Gal}(K:F) = \text{Gal}(g, F)$ is soluble, and Proposition 7.39 says that a quotient of a soluble group is itself soluble, so $\text{Gal}(f, F) = \text{Gal}(E:F)$ is soluble as claimed.

Now suppose $k > 1$. Again, let $a = \alpha_1^{n_1}$, and let K be the splitting field of $g = x^{n_1} - a$ over F . Let L be the splitting field of g over E . Then $F \subseteq E \subseteq K \subseteq L$. Furthermore, L is the splitting field of gf over F , and the splitting field of f over K . Since $F(\alpha_1) \subseteq K$ it follows that f splits in $K(\alpha_2, \dots, \alpha_k)$.

That is, since $f \in F[x]$ and $F \subseteq K$, we can regard f as an element of $K[x]$. Its splitting field over K is contained in $K(\alpha_2, \dots, \alpha_k)$, a radical extension of K by $(k-1)$ adjoined elements. By the induction hypothesis, this means that $\text{Gal}(f, K) = \text{Gal}(L : K)$ is soluble.

By Proposition 11.110, $\text{Gal}(g, F) = \text{Gal}(K : F)$ is soluble. By part (iv) of the Fundamental Theorem of Galois Theory,

$$\text{Gal}(g, F) = \text{Gal}(K : F) \cong \text{Gal}(L : F) / \text{Gal}(L : K)$$

and since $\text{Gal}(L : K)$ is soluble, Proposition 7.40 says that $\text{Gal}(L : F)$ is soluble, as an extension of soluble groups.

Then by the Fundamental Theorem of Galois Theory again,

$$\text{Gal}(f, F) = \text{Gal}(E : F) \cong \text{Gal}(L : F) / \text{Gal}(L : E)$$

and by Proposition 7.39 this is soluble as a quotient of a soluble group.

This completes the induction step, and thereby the proof. $\square$

Combining Propositions 11.102 and 11.109 we get the following:

Theorem 11.111 *Let F be a field of characteristic zero. Then a polynomial $f \in F[x]$ is soluble by radicals if and only if $\text{Gal}(f, F)$ is soluble.*

The case of nonzero characteristic is rather more complicated, and we will state it without proof:

Theorem 11.112 *Let F be a field of prime characteristic p , and suppose that $f \in F[x]$ has splitting field E over F .*

If $\text{Gal}(f, F) = \text{Gal}(E : F)$ is soluble and has order not divisible by p , then f is soluble by radicals.

Conversely, if $E : F$ is a radical extension with intermediate subfields

$$F = F_0 \subset F_1 \subset \cdots \subset F_k = E$$

with $F_i = F_{i-1}(\alpha_i)$ for a root α_i of a polynomial $g_i = x^{n_i} - a_i \in F_{i-1}[x]$, and if $n_1 n_2 \cdots n_k$ is not divisible by p , then $\text{Gal}(f, F) = \text{Gal}(E : F)$ is soluble.

This means that polynomials with insoluble Galois groups aren't soluble by radicals. The smallest insoluble groups we met in Section 7.3 were the alternating group A_5 and the symmetric group S_5 . The following result guarantees the existence of polynomials with Galois group isomorphic to S_p for any prime p .

Proposition 11.113 *Let $f \in \mathbb{Q}[x]$ be a monic, irreducible polynomial of prime degree p , with exactly two roots in $\mathbb{C} \setminus \mathbb{R}$. Then the Galois group $\text{Gal}(f, \mathbb{Q}) \cong S_p$.*

Proof Let E be the splitting field of f over $\mathbb{Q}$; this is contained in $\mathbb{C}$. By Proposition 11.84, f is separable over $\mathbb{Q}$, and hence has p distinct roots. The Galois group $\text{Gal}(f, \mathbb{Q}) = \text{Gal}(E : \mathbb{Q})$ is a group of permutations of these roots, and is therefore a subgroup of S_p .

Let α be a root of f . Then f is the minimal polynomial of the extension $\mathbb{Q}(\alpha) : \mathbb{Q}$, and by Proposition 11.30 and Corollary 11.95 we have

$$|\text{Gal}(\mathbb{Q}(\alpha) : \mathbb{Q})| = [\mathbb{Q}(\alpha) : \mathbb{Q}] = \deg(f) = p.$$

Applying part (ii) of the Fundamental Theorem of Galois Theory,²⁹

²⁹ Theorem 11.98, page 463.

$$|\text{Gal}(E : \mathbb{Q})| / |\text{Gal}(E : \mathbb{Q}(\alpha))| = |\text{Gal}(\mathbb{Q}(\alpha) : \mathbb{Q})| = p,$$

so the order of $\text{Gal}(f, \mathbb{Q}) = \text{Gal}(E : \mathbb{Q})$ is divisible by p . By Cauchy's Theorem³⁰ $\text{Gal}(E : \mathbb{Q})$ thus has an element of order p . But the only elements of order p in S_p are p -cycles, so $\text{Gal}(E : \mathbb{Q})$ contains such a permutation. The conjugation map swaps the two complex roots of f and leaves the real roots fixed, and thereby determines a transposition in S_p . By Proposition 1.60 (iii), these two permutations generate all of S_p , and hence $\text{Gal}(f, \mathbb{Q}) = \text{Gal}(E : \mathbb{Q}) \cong S_p$. $\square$

³⁰ Theorem 2.37, page 57.

We can illustrate this with a simple example. The argument relies on two fundamental results from real analysis, the Intermediate Value Theorem³¹ and Rolle's Theorem.³²

³¹ Theorem A.41, page 502.

³² Theorem A.42, page 502.

Example 11.114 Let $f = x^5 - 5x + 2 \in \mathbb{Q}[x]$. This is monic, and irreducible over $\mathbb{Q}$ by Eisenstein's Criterion. By evaluating f at a few test values, we find that

$$f(-2) = -20, \quad f(-1) = 6, \quad f(0) = 2, \quad f(1) = -2, \quad f(2) = 24.$$

By the Intermediate Value Theorem, the intervals $(-2, -1)$, $(0, 1)$ and $(1, 2)$ thus contain at least one root each.

We can use Rolle's Theorem to show that each of these intervals contains exactly one root. The first derivative $f' = 5x^4 - 5$ is zero at $x = \pm 1$, and nowhere else. If there were more than one root in $(-1, 1)$ then Rolle's Theorem would imply the existence of another turning point in this interval, which can't happen.

Similarly, $f'(x)$ is strictly positive for $|x| > 1$, so again f can have at most one root in each of the intervals $(-\infty, -1)$ and $(1, \infty)$. Putting all this together tells us that f has exactly three real roots, one in each of the intervals $(-2, -1)$, $(0, 1)$ and $(1, 2)$.

The remaining two roots must thus be complex, and hence by Proposition 11.113 the Galois group $\text{Gal}(f, \mathbb{Q})$ is isomorphic to the symmetric group S_5 . This isn't soluble by Proposition 7.37, and so f isn't soluble by radicals by Proposition 11.109.

This doesn't mean that f has no roots, nor that it doesn't have a complete set of roots in some appropriate algebraic extension field. All it means is that the roots of f can't be expressed using just the basic arithmetical operations and finite roots.

We can certainly use numerical methods (such as the Newton-Raphson

iterative method) to calculate the roots of f to arbitrary accuracy, but this is algebraically unsatisfying. Another approach is to extend the class of operations we allow: following an approach pioneered by the Swedish mathematician Erland Bring (1736–1798) and the British mathematician George Jerrard (1804–1863), we define the **ultraradical** or **Bring radical** $B(a)$ of $a \in \mathbb{Q}$ to be the real root of $x^5 + x + a$. By a suitable transformation, any quintic polynomial can be reduced to the form $x^5 + px + q$ (the **Bring–Jerrard normal form**), whose roots are

$$\sqrt[4]{-\frac{p}{5}} B\left(-\frac{1}{4}\left(-\frac{5}{p}\right)^{5/4} q\right)$$

and its four conjugates.

Charles Hermite (1822–1901) discovered a general solution for the quintic equation in terms of elliptic functions, and later Felix Klein (1849–1925) developed a general theory involving elliptic functions and the rotational symmetry group of the regular icosahedron (which is isomorphic to the alternating group A_5).

The Academy has taken, this year, the decision to no longer examine any solution of the problems of the doubling of the cube, of the trisection of the angle, or of the squaring of the circle, nor any announcement of a perpetual motion machine.

— *Histoire de l'Académie Royale des Sciences* (1775) 61

I consider myself as having made my report, and being discharged from further attendance on the subject. I will not, from henceforward, talk to any squarer of the circle, trisector of the angle, duplicator of the cube, constructor of perpetual motion, subverter of gravitation, stagnator of the earth, builder of the universe, &c.

— Augustus De Morgan (1806–1871),
A Budget of Paradoxes (1872) 7

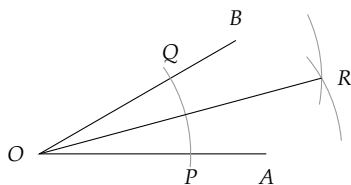


Figure 11.5: Angle bisection with compass and straightedge

11.A Geometric constructions

IN THIS SECTION we will look at some applications of Galois Theory to classical geometry. In particular, we will identify what geometric constructions are possible using just an unmarked ruler and a pair of compasses. As ancient Greek mathematicians discovered, we can achieve a surprisingly wide range of geometric tasks with just these tools.

Example 11.115 (Bisecting an angle) Given an arbitrary angle θ , we can construct an angle $\frac{\theta}{2}$ with straightedge and compasses as shown in Figure 11.5.

Set the compasses to some arbitrary distance d , place the point at O and sketch out an arc of radius d . This arc intersects the lines OA and OB at points P and Q respectively. Now put the compass point at P and sketch out an arc of some radius r (where r may or may not equal d , it doesn't really matter). Then do the same at Q . These two arcs intersect at a point R . The line OR intersects the angle AOB ; that is, the angles AOR and ROB are both equal to $\frac{\theta}{2}$.

There are three famous, unsolved classical problems in geometry that between them occupied mathematicians for countless hours over at least two and a half thousand years and possibly longer. Each of these must be achieved, for any arbitrary circle, angle or cube, just using a pair of compasses and an unmarked ruler.

Squaring the circle Construct a square with exactly the same area as a given circle.

Trisecting an angle Construct an angle exactly equal to one third of a given angle.

Doubling a cube Construct a cube with volume exactly twice that of a given cube.

In the middle of the 19th century, all three of these were finally proved to be impossible,³³ and we will study the reasons why now.

We'll do this by developing a characterisation of those points in the plane $\mathbb{R}^2$ that can be obtained just by using an unmarked ruler and a pair of compasses, and then using this to construct a hierarchy of field extensions of $\mathbb{Q}$.

First we choose a coordinate system for $\mathbb{R}^2$: an origin O and two linearly independent vectors $\mathbf{e}_1$ and $\mathbf{e}_2$, which determine the coordinate axes. Then any other vector $\mathbf{v}$ in $\mathbb{R}^2$ can be expressed uniquely as a linear combination $\mathbf{v} = a\mathbf{e}_1 + b\mathbf{e}_2$, and this corresponds to the point with coordinates (a, b) .

Suppose that P is a set of points in $\mathbb{R}^2$. We consider the following two operations:

- (L) Draw a line through any two distinct points in P .
- (C) Draw a circle centred on some point in P , with radius equal to the distance between any two distinct points in P .

Let $C(P)$ be the set of points in $\mathbb{R}^2$ that can be obtained as the intersection of two lines, two circles, or a line and a circle, resulting from applying operations (L) and (C) to the points in P . The points in $C(P)$ are said to be **constructible in one step** from P .

Now let $P_0 = \{(0, 0), (1, 0)\}$. We recursively define a sequence of sets $P_0, P_1, P_2, \dots$ of points in $\mathbb{R}^2$ such that

$$P_i = P_{i-1} \cup C(P_{i-1}).$$

That is, each set P_i consists of the points in P_{i-1} together with all the points which are constructible in one step from P_{i-1} . We say any point in $\mathbb{R}^2$ is **constructible** if it lies in P_n for some n ; that is, if it is constructible in finitely many steps from P_0 .

The set P_1 is constructed from P_0 by means of the operations (L) and (C). The only possible line at this stage is the horizontal axis, passing through $(0, 0)$ and $(1, 0)$. There are two possible circles that can be drawn, one centred at $(0, 0)$ and the other at $(1, 0)$, both with radius 1. These intersect at the points $(\frac{1}{2}, \pm \frac{\sqrt{3}}{2})$, so

$$P_1 = \{(0, 0), (1, 0), (\frac{1}{2}, \frac{\sqrt{3}}{2}), (\frac{1}{2}, -\frac{\sqrt{3}}{2})\}.$$

The circles and lines passing through these four points yield several

³³ Impossible with just straightedge and compasses, that is: the latter two are possible using slightly more sophisticated techniques. Solutions involving the intersection of conic curves or marked rulers were known to Greek mathematicians in the fourth century BCE. More recently, solutions have been devised using origami, linkages and similar mechanisms.

more, the elements of P_2 , and by continuing this process indefinitely. To translate this into an algebraic context and use the machinery developed in this and preceding chapters, we define a tower of fields

$$F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n \subseteq \cdots \subseteq \mathbb{R}$$

by setting each F_n to be the subfield of $\mathbb{R}$ generated by the coordinates of the points in P_n . For each point (x, y) in $P_n \setminus P_{n-1}$ we adjoin both x and y to the field F_{n-1} . So F_0 is the field generated by the elements 0 and 1, the coordinates of the points in $P_0 = \{(0, 0), (1, 0)\}$; thus $F_0 = \mathbb{Q}$. To get F_1 we adjoin the elements $\frac{1}{2}$ and $\frac{\sqrt{3}}{2}$, the coordinates of the points in $P_1 \setminus P_0$, which yields $F_1 = \mathbb{Q}(\sqrt{3})$.

This new field F_1 is an algebraic extension of $F_0 = \mathbb{Q}$ with degree $[\mathbb{Q}(\sqrt{3}) : \mathbb{Q}] = 2$. In fact, each successive extension $F_n : F_{n-1}$ (and, by the Tower Law,³⁴ any extension $F_n : F_m$) has degree 2^k for some $k \in \mathbb{N}$:

³⁴ Proposition 11.21, page 428.

Proposition 11.116 *Let F_n be the subfield of $\mathbb{R}$ generated by the coordinates of the constructible points in the set P_n discussed above. Then $[F_n : \mathbb{Q}]$ is a power of 2 for all $n \in \mathbb{N}$.*

Proof We proceed by induction on n . As discussed earlier, $[F_0 : \mathbb{Q}] = 1 = 2^0$ and $[F_1 : \mathbb{Q}] = 2$.

Now suppose that $[F_{n-1} : \mathbb{Q}] = 2^k$ for some $k \in \mathbb{N}$. Each of the points in $P_n \setminus P_{n-1}$ is an intersection of two lines, a line and a circle, or two circles determined by points in P_{n-1} . We consider each of these cases in turn.

Case 1 Suppose we have four points $A = (a_1, a_2)$, $B = (b_1, b_2)$, $C = (c_1, c_2)$ and $D = (d_1, d_2)$. The equation of the lines AB and CD are

$$\frac{x - a_1}{b_1 - a_1} = \frac{y - a_2}{b_2 - a_2}, \quad \frac{x - c_1}{d_1 - c_1} = \frac{y - c_2}{d_2 - c_2}.$$

Finding the intersection point of AB and CD amounts to solving these two simultaneous linear equations, a task which requires only addition, subtraction, multiplication and division. Hence both x and y already lie in F_{n-1} , so $F_{n-1}(x, y) = F_{n-1}$ and thus $[F_{n-1}(x, y) : F_{n-1}] = 1$.

Case 2 Given three points $A = (a_1, a_2)$, $B = (b_1, b_2)$ and $C = (c_1, c_2)$, we now consider the intersection of the line AB with the circle centred at C with radius $r \in F_{n-1}$. The equation of the line AB is

$$\frac{x - a_1}{b_1 - a_1} = \frac{y - a_2}{b_2 - a_2}$$

and the equation of the circle centred at C is

$$(x - c_1)^2 + (y - c_2)^2 = r^2.$$

Solving these we get

$$(x - c_1)^2 + \left(\frac{b_2 - a_2}{b_1 - a_1} (x - a_1) + a_2 - c_2 \right)^2 = r^2$$

$$\left(\frac{b_1 - a_1}{b_2 - a_2} (y - a_2) + a_1 - c_1 \right)^2 + (y - c_2)^2 = r^2$$

so x and y are roots of quadratic polynomials over F_{n-1} . Then $[F_{n-1}(x, y) : F_{n-1}] = 1, 2$ or 4 , depending on whether both, only one or neither of x and y lie in F_{n-1} .

Case 3 The third case to consider is the intersection of two circles. Given circles centred at $A = (a_1, a_2)$ and $B = (b_1, b_2)$ with radii r and s respectively, where $r, s \in F_{n-1}$, we get equations

$$(x - a_1)^2 + (y - a_2)^2 = r^2, \quad (x - b_1)^2 + (y - b_2)^2 = s^2.$$

Expanding these and subtracting one from the other yields

$$2(b_1 - a_1)x + 2(b_2 - a_2)y + a_1^2 + a_2^2 - b_1^2 - b_2^2 = r^2 - s^2,$$

a linear equation in x and y , which determines the chord between the two intersection points. This case thus reduces to the previous case of a line intersecting with a circle, and so $[F_{n-1}(x, y) : F_{n-1}] = 1, 2$ or 4 again.

In each case, then, adjoining elements determined by these intersection points thus multiplies the degree of the extension by $1, 2$ or 4 . So adjoining to F_{n-1} elements corresponding to all the points in $P_n \setminus P_{n-1}$ results in the degree $[F_n : F_{n-1}] = 2^j$ for some non-negative integer j .

By the Tower Law,³⁵ $[F_n : \mathbb{Q}] = [F_n : F_{n-1}][F_{n-1} : \mathbb{Q}] = 2^j \times 2^k = 2^{j+k}$, ³⁵ Proposition 11.21, page 428. which completes the induction step and thereby the proof. $\square$

We can now apply this to the three classical geometric problems mentioned earlier. The impossibility of two of these was first proved by the French mathematician Pierre Wantzel (1814–1848).

Corollary 11.117 *A cube cannot be doubled in volume using just ruler and compasses.*

Proof Suppose, without loss of generality, that the cube in question has sides of unit length, and hence unit volume. A cube of twice this volume must have sides of length α where $\alpha^3 = 2$, whose minimal polynomial is $x^3 - 2 \in \mathbb{Q}[x]$. By Eisenstein's Criterion³⁶ this is irreducible over $\mathbb{Q}$, and hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$. But by Proposition 11.116 constructible lengths and points must lie in field extensions of degree 2^k for $k \in \mathbb{N}$, and hence no such construction exists for α . $\square$

³⁶ Proposition 11.1, page 422.

Corollary 11.118 *Not all angles can be trisected using just ruler and compasses.*

Proof As a counterexample, let $\theta = \frac{\pi}{3}$. To trisect this angle requires us to construct $\phi = \theta/3 = \frac{\pi}{9}$, or equivalently the length $y = \cos \phi$. Using the trigonometric identity

$$\cos 3\phi = 4\cos^3 \phi - 3\cos \phi$$

and the fact that $\cos 3\phi = \cos \frac{\pi}{3} = \frac{1}{2}$ this gives

$$8y^3 - 6y - 1 = 0.$$

Setting $x = 2y$ this yields

$$x^3 - 3x - 1 = 0$$

and so this particular trisection problem is equivalent to constructing a length α satisfying this equation. But the polynomial $x^3 - 3x - 1$ is irreducible over $\mathbb{Q}$, again by Eisenstein's Criterion, the extension $\mathbb{Q}(\alpha) : \mathbb{Q}$ has degree 3. $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$, which is not divisible by 2. By Proposition 11.116 no such construction exists. $\square$

Some angles can be trisected in this way, however. If $\theta = \frac{\pi}{2}$ then the question concerns the polynomial $4x^3 - 3x$ which factorises as $x(4x^2 - 3)$ and so the corresponding field extension is $\mathbb{Q}(\sqrt{3}) : \mathbb{Q}$. This has degree 2, and $\frac{\pi}{2}$ can be trisected in the required manner.

The other classical problem, that of squaring the circle, requires the fact that π is not algebraic over $\mathbb{Q}$, a fact originally proved in 1882 by the German mathematician Ferdinand Lindemann (1852–1939).

Corollary 11.119 *A circle cannot be squared using just ruler and compasses.*

Proof Without loss of generality, consider a circle of unit radius. This has area π units, and so a square of the same area must have side $\sqrt{\pi}$. To construct such a square requires the point $(0, \sqrt{\pi})$ to be constructible from $P_0 = \{(0, 0), (1, 0)\}$ in finitely many steps. By Proposition 11.116 this requires the extension $\mathbb{Q}(\sqrt{\pi}) : \mathbb{Q}$ to be algebraic with degree a power of 2. But $\mathbb{Q}(\pi) \subseteq \mathbb{Q}(\sqrt{\pi})$, and π is known not to be algebraic over $\mathbb{Q}$, so no such construction exists. $\square$

Summary

References and further reading

In this chapter we've only been able to look at the main ideas of Galois theory. The following two books, both very readable and aimed at undergraduate students, cover the subject more thoroughly:

- J M Howie, *Fields and Galois Theory*, Springer Undergraduate Mathematics Series, Springer (2006)
- I N Stewart, *Galois Theory*, fourth edition, CRC Press (2015)

The latter also includes a couple of chapters of historical and biographical notes. The following article gives a much briefer overview of the subject:

- J Stillwell, *Galois Theory for Beginners*, The American Mathematical Monthly 101.1 (1994) 22–27

The romantic and tragic story of Galois' brilliant but short life has passed into mathematical folklore and, as often happens, many of the things we think we know are wrong (particularly the popular image of the doomed mathematician frantically writing down his discoveries in the hours before his fatal duel). The following article unpicks many of these myths, giving an account of Galois' life and death better supported by the available historical sources:

- T Rothman, *Genius and Biographers: The Fictionalization of Évariste Galois*, The American Mathematical Monthly 89.2 (1982) 84–106

The next two articles present applications of finite fields to mathematical puzzles.

- N G de Bruijn, *A solitaire game and its relation to a finite field*, Journal of Recreational Mathematics 5.2 (1972) 133–137
- D A James, *Magic circles*, Mathematics Magazine 54.3 (1981) 122–125

Exercises

- 12.1 Use Proposition 11.5 together with the prime $p = 29$, in a manner analogous to the method used in Example 11.11, to find a list of quadratic and higher-degree irreducible polynomials in $\mathbb{Z}[x]$.
- 12.2 Show that the extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is simple.
- 12.3 Adapt the argument in Example 11.82 to show that if $[E : F] = 2$ and $\text{char}(F) \neq 2$, then $E : F$ is separable. What happens if $\text{char}(F) = 2$?

We have offended against thy holy lawes: We have left undone those things whiche we ought to have done, and we have done those things which we ought not to have done, and there is no health in us, but thou, O Lorde, have mercy upon us miserable offendours.

— The Book of Common Prayer (1559)

A Background

THE REST OF THIS BOOK assumes some background knowledge that is usually taught early in the first year of a mathematics degree programme at a British university.

This appendix comprises a quick overview of some of the more relevant bits of this material, in order to make this book accessible to a wider audience, to provide a quick reminder for those who might need one, and also to standardise notation and terminology. In particular, we cover sets, infinite cardinals, functions, and relations.

Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns: the ones we don't know we don't know.

— Donald Rumsfeld, US Department of Defense news briefing, 12 February 2002

A.1 Sets

MUCH OF PURE MATHEMATICS, and abstract algebra in particular, is based on set theory in one form or another; indeed, almost all of the objects we study in this book are essentially sets equipped with some extra structure.

Much has been written over the years, particularly during the earlier and middle years of the 20th century, on axiomatic set theory: the formalisation and underpinnings of much of the foundations of pure mathematics. This is in many ways an interesting subject, and some quite surprising results and unexpectedly tricky problems arise in the process, but it is beyond the scope of the current discussion. We will adopt a more concrete approach (sometimes called **naïve set theory**), and merely use the machinery without worrying too much about how it works, and in what circumstances it might break down. Although awkward paradoxes can occur, often when we consider sets of sets without due care and attention, in practice we won't run into any of these situations in the rest of this book.

Broadly speaking, a **set** is an amorphous collection of objects, a bag of distinct things. Although most of the more familiar examples will tend to have some additional structure, such as an ordering, an addition or

A set is a many that allows itself to be thought of as a One.

— Georg Cantor (1845–1918)



Wikimedia Commons / Sophia De Morgan (1809–1892)

Augustus De Morgan (1806–1871) was born in Madras, but when he was ten his father (also called Augustus) died and the family returned to England. His mother, a devout Anglican, hoped he would become a clergyman but at sixteen he entered Trinity College, Cambridge to study mathematics. He played the flute and enjoyed learning for its own sake, which hampered his performance in the fiercely competitive Cambridge mathematics tripos exams, and he graduated fourth in his class in 1827, with the title of Fourth Wrangler and the degree of Bachelor of Arts.

His nonconformist beliefs prevented him from taking the theological exam required for progression to the degree of Master of Arts and a college fellowship, so he left Cambridge and briefly studied law at Lincoln's Inn, but in 1828 was appointed Professor of Mathematics at the newly-founded London University. He stayed there until 1831, when he and several colleagues resigned *en masse* on a point of principle. His successor drowned in 1836 and De Morgan was reappointed, remaining until 1866, when he resigned again on a point of principle.

In addition to his many contributions to logic, set theory and algebra, in 1865 De Morgan and his son George founded the London Mathematical Society, which to this day is the main British learned society for pure and applied research mathematicians. He was also an entertaining writer: his book *A Budget of Paradoxes*, published posthumously in 1872, is a fascinating study of mathematical crackpottery.

¹ This symbol $\emptyset$ is not to be confused with the lowercase Greek letter ϕ or φ .

multiplication operation, or some notion of distance between elements, by default a set has no preordained or canonical internal structure.

Definition A.1 A **set** is a (finite or infinite) collection of distinct objects, which are called the **elements** or **members** of the set.

We denote membership by the symbol $\in$, so $x \in S$ means that x is an element of the set S , while $y \notin S$ means that y is not an element of S .

We will usually adopt the convention that sets are represented by capital letters, while elements are denoted by lower-case letters.

There are various ways of defining a given set. One, suitable only for finite sets (or very occasional instances of particularly well-behaved infinite sets) is to simply list the elements:

$$A = \{a, b, c, d\},$$

$$B = \{\pm 1, \pm\sqrt{2}, \pm\sqrt{3}, \dots\},$$

$$C = \{\dots, -4, -2, 0, 2, 4, \dots\},$$

$$D = \{\text{Io, Europa, Ganymede, Callisto}\}.$$

Sometimes it's more convenient to specify the elements in some other, well-defined way:

$$S^1 = \{(x, y) : x, y \in \mathbb{R}, x^2 + y^2 = 1\}$$

This set consists of all ordered pairs of real numbers (x, y) such that $x^2 + y^2 = 1$; it can be realised geometrically as the unit circle in the Euclidean plane $\mathbb{R}^2$.

In all of these examples we have used braces $\{$ and $\}$ to contain the elements or their specifications. In the last example, the colon $:$ should be interpreted as “such that”, so we could read the whole thing as “ S^1 is the set of all (x, y) such that x and y are in $\mathbb{R}$ and $x^2 + y^2 = 1$ ”. Some books use a vertical bar $|$ instead of a colon.

Example A.2 Some standard sets that we refer to regularly throughout the rest of the book include the following:

$\mathbb{N} = \{1, 2, 3, \dots\}$	the natural numbers
$\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$	the integers
$\mathbb{Q} = \{\frac{a}{b} : a, b \in \mathbb{Z}, b > 0, \gcd(a, b) = 1\}$	the rational numbers
$\mathbb{R}$	the real numbers
$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}, i^2 = -1\}$	the complex numbers

One particularly (if unexpectedly) important set is the **empty set**, the set with no elements. We denote it $\emptyset = \{\}$.¹ Also important, but generally only in a formal sense, is the **universal set** $\mathcal{E}$. This can be regarded as the set containing everything (including every other set); however be warned that one can easily run into awkward paradoxes

if one isn't careful, especially when dealing with sets of sets.²

Because of this, many versions of axiomatic set theory (including **Zermelo–Fraenkel set theory**) don't permit the existence of a universal set. To sidestep these problems, we will typically use $\mathcal{E}$ to denote the largest set necessary in a particular context. This is a somewhat imprecise way of doing things, but will suffice for our purposes.

The sets in Example A.2 can all be nested neatly inside each other: the natural numbers are the positive integers, the integers are the rational numbers with denominator 1, the rationals are those real numbers that can be expressed as a quotient, and the real numbers are the complex numbers with imaginary part equal to zero.

Definition A.3 Let A and B be two sets. We say that A is a **subset** of B , written $A \subseteq B$ or $B \supseteq A$, if every element of A is an element of B as well.

If B also contains at least one extra element that isn't in A , then we say A is a **proper subset** of B , and write $A \subset B$ or $B \supset A$.

In particular, the empty set $\emptyset$ is a subset of every set, including itself: all of the no elements $\emptyset$ doesn't have are also contained in any other set we care to consider. Any set contains itself as a subset (although obviously not a proper subset), and our somewhat nebulously-defined universal set $\mathcal{E}$ contains itself and every other set under consideration.

Many important theorems in real analysis, such as the **Intermediate Value Theorem** or **Taylor's Theorem**, are concerned with the continuity or differentiability of functions over contiguous subsets of $\mathbb{R}$, and so it is useful to have a compact notation for this situation.

Definition A.4 Let $a, b \in \mathbb{R}$ such that $a < b$. Then the **closed interval** $[a, b]$ is the set

$$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$$

and the **open interval** (a, b) is the set

$$(a, b) = \{x \in \mathbb{R} : a < x < b\}.$$

We may also define the **half-open intervals**

$$[a, b) = \{x \in \mathbb{R} : a \leq x < b\}, \quad (a, b] = \{x \in \mathbb{R} : a < x \leq b\}.$$

A variant notation used by some books replaces parentheses with reversed square brackets, for example writing the open interval (a, b) as $]a, b[$. We can also extend this notation to infinite endpoints:

$$\begin{aligned} (a, \infty) &= \{x \in \mathbb{R} : x > a\}, & [a, \infty) &= \{x \in \mathbb{R} : x \geq a\}, \\ (-\infty, b) &= \{x \in \mathbb{R} : x < b\}, & (-\infty, b] &= \{x \in \mathbb{R} : x \leq b\}. \end{aligned}$$

We will often want to combine or compare two or more sets, and this

²One such paradox is sometimes known as **Russell's Paradox** (after the philosopher and mathematician Bertrand Russell (1872–1970)) or the **Barber's Paradox**. The latter form is usually stated as follows:

Suppose there exists a town whose residents are extremely pogonophobic; that is, they are either afraid of, strongly disapprove of, or are otherwise prejudiced against men with beards. The town has exactly one professional barber, who shaves every man who doesn't shave himself. The seemingly innocuous question "who shaves the barber?" leads to an irresolvable inconsistency: if the barber (in his capacity as a private citizen) doesn't shave himself, then he must do so (in his capacity as the town's barber); if on the other hand he does, then he shouldn't.

There are other ways of formulating this paradox: by considering a book indexing all books that don't cite themselves, or by contemplating a set that contains every set that doesn't contain itself.

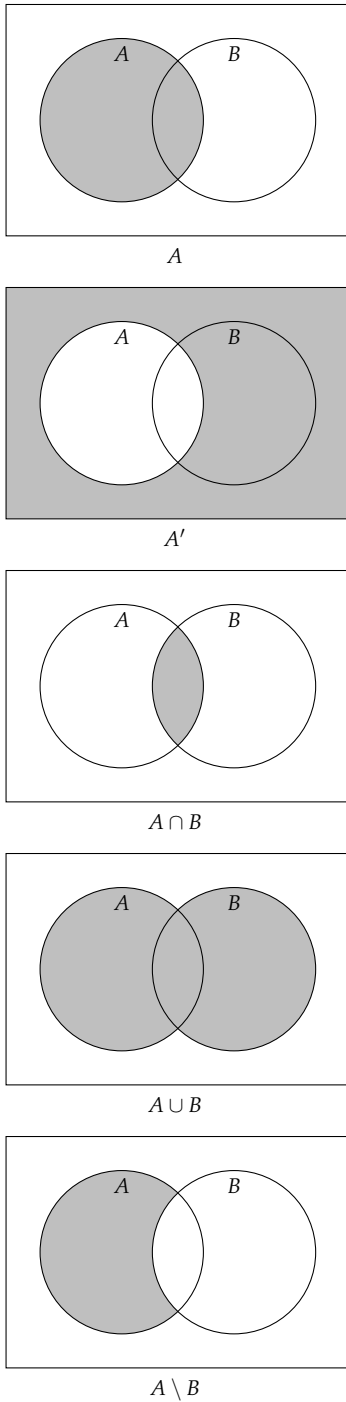


Figure A.1: Venn diagrams depicting standard set operations

leads us to the following two operations.

Definition A.5 The **union** $A \cup B$ of two sets A and B is defined as

$$A \cup B = \{x : x \in A \text{ and/or } x \in B\}.$$

It is the set formed by combining the elements of A and B into one larger collection and discarding any duplicates.

The **intersection** of A and B is

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

and comprises those elements that are in both A and B .

If $A \cap B = \emptyset$ then we say that A and B are **disjoint**.

We can extend this to arbitrary finite collections of sets. We need to check that both the union and intersection operations can be performed in any order without changing the end result; that is,

$$(A \cup B) \cup C = A \cup (B \cup C) \quad \text{and} \quad (A \cap B) \cap C = A \cap (B \cap C)$$

for any sets A, B and C . We say that $\cup$ and $\cap$ are **associative** operations. Similarly, with a bit of thought we can see that it doesn't matter whether we consider $A \cup B$ or $B \cup A$, because we'll get the same result either way. The same is true for intersections: $A \cap B = B \cap A$. (Operations satisfying this order-invariance condition are said to be **commutative**.)

Since $\cup$ and $\cap$ are associative, we can ignore parentheses in unions and intersections of finite collections of sets. This means that, given a collection $A_1, \dots, A_n$ of sets we can define

$$\bigcup_{i=1}^n A_i = A_1 \cup \dots \cup A_n \quad \text{and} \quad \bigcap_{i=1}^n A_i = A_1 \cap \dots \cap A_n$$

secure in the knowledge that we'll get the same answer whatever order we do each pairwise union or intersection operation.

Another important operation is the **difference** of two sets, which is defined as

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

If A and B are disjoint then $A \setminus B = A$. In general, it is not the case that $A \setminus B = B \setminus A$, so the difference operation is **noncommutative**.

The **complement** of a set A is

$$A' = \mathcal{E} \setminus A = \{x : x \notin A\},$$

and consists of all elements that aren't in A .

But what do we really mean by this? We need some well-defined way of saying when two sets are equal, and the answer is the obvious one: two sets are equal if they contain exactly the same elements.

Table A.1 contains a standard list of general laws or identities satisfied by the set operations we've met so far.

Commutative Laws	$\begin{cases} A \cup B = B \cup A \\ A \cap B = B \cap A \end{cases}$	Associative Laws	$\begin{cases} A \cup (B \cap C) = (A \cup B) \cap C \\ A \cap (B \cup C) = (A \cap B) \cup C \end{cases}$
Distributive Laws	$\begin{cases} A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \end{cases}$	Identity Laws	$\begin{cases} A \cup \emptyset = A \\ A \cap \mathcal{E} = A \end{cases}$
Complement Laws	$\begin{cases} A \cup A' = \mathcal{E} \\ A \cap A' = \emptyset \\ (A')' = A \end{cases}$	Absorption Laws	$\begin{cases} A \cup (A \cap B) = A \\ A \cap (A \cup B) = A \end{cases}$
Minimisation Laws	$\begin{cases} (A \cap B) \cup (A \cap B') = A \\ (A \cup B) \cap (A \cup B') = A \end{cases}$	De Morgan's Laws	$\begin{cases} (A \cup B)' = A' \cap B' \\ (A \cap B)' = A' \cup B' \end{cases}$

Table A.1: Standard identities in set algebra

We can depict these geometrically by means of **Venn diagrams**. These consist of a boundary rectangle, representing $\mathcal{E}$, and inner circles (or other closed curves) representing other sets. We represent intersections by shading in the overlapping portions of two or more circles, and unions by shading in all of the sets in question.

We can check the identities in Table A.1 by drawing their Venn diagrams; for example, both sides of the distributive law

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

yield the diagram in Figure A.2.

As useful as they are for illustrative purposes, Venn diagrams don't constitute a formal proof, so we need a better method of proving set equality. As remarked earlier, two sets are equal if they have exactly the same elements. Equivalently, for A and B to be equal, every element in A must be in B , and every element in B must also be in A . This is the same as requiring A to be a subset of B and B to be a subset of A at the same time. So, $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$.

To do this, we consider some arbitrary element of A and show that it must also be in B . Then we consider some arbitrary element of B and show that it must be in A . The following proof illustrates this.

Proposition A.6 *Let A , B and C be sets. Then*

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

Proof Suppose that $x \in A \cup (B \cap C)$. Then either $x \in A$ or $x \in B \cap C$ (or both). We'll consider both of these separately.

If $x \in A$ then it's certainly the case that $x \in A \cup B$ and $x \in A \cup C$, because each of these sets are just A with the elements of, respectively, B or C included. And since x lies in both $A \cup B$ and $A \cup C$, then x

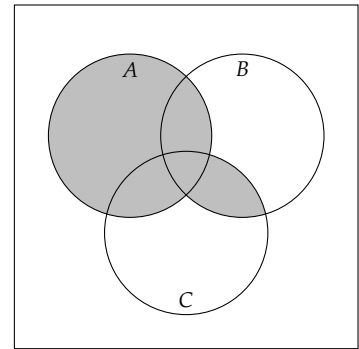
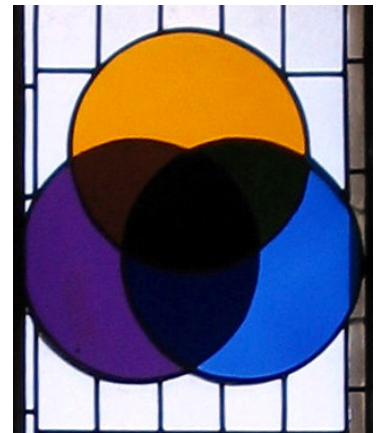


Figure A.2: Venn diagram depicting the distributive law

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$



Wikimedia Commons / User:Schutz

Figure A.3: Detail from a memorial window to John Venn (1834–1923) in the dining hall of Gonville and Caius College, Cambridge



Wikimedia Commons

John Venn (1834–1923) was born in Yorkshire, a descendent of a line of influential Anglican clergymen who campaigned strongly on social issues, such as prison reform, the abolition of slavery, and banning cruel sports.

John himself studied mathematics at Gonville and Caius College, Cambridge, graduating as Bachelor of Arts in 1857 with the sixth highest mark in his year, winning the title of Sixth Wrangler and a college fellowship. Following in the family tradition he entered the Church of England, being ordained first as a deacon in 1858 and as a priest a year later. He served as Curate of Mortlake in Surrey until 1862, when he returned to Cambridge as a College Lecturer in Moral Sciences.

In addition to his work on logic and set theory, he wrote several books on the history of the University of Cambridge, in particular the ten-volume *Alumni Cantabrigienses*, compiled in collaboration with his son, the economist John Archibald Venn (1883–1958). This comprises detailed biographical notes on all known graduates of the University from its foundation in the early 13th century until the mid 20th century.

He was elected a Fellow of the Royal Society in 1883, granted the degree of Doctor of Science in 1884, elected a Fellow of the Society of Antiquaries in 1892, and was President of Gonville and Caius College from 1912 until 1923. He was also a keen amateur engineer, his most successful creation being a cricket bowling machine that bowled out some members of the Australian national team during their visit to Cambridge in 1909.

must also be in their intersection. Hence $x \in (A \cup B) \cap (A \cup C)$.

Alternatively, if $x \in B \cap C$, then that means that $x \in B$ and $x \in C$. Hence x is a member of both $A \cup B$ and $A \cup C$, and must therefore be in their intersection. Thus $x \in (A \cup B) \cap (A \cup C)$, and so $A \cup (B \cap C) \subseteq (A \cup B) \cap (A \cup C)$.

Now suppose $x \in (A \cup B) \cap (A \cup C)$. Then x is a member of both $A \cup B$ and $A \cup C$. Again, we have two cases: either $x \in A$ or $x \notin A$.

If $x \in A$ then it must also be in $A \cup (B \cap C)$, since this is just A with the elements of $B \cap C$ included as well.

If, however, $x \notin A$ then it must lie in the other parts of $A \cup B$ and $A \cup C$, which means that $x \in B$ and $x \in C$. Hence $x \in B \cap C$, and therefore $x \in A \cup (B \cap C)$.

Thus $(A \cup B) \cap (A \cup C) \subseteq A \cup (B \cap C)$, and therefore $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ as claimed. $\square$

Another important set operation that we use regularly throughout the rest of the book is the **Cartesian product**:

Definition A.7 Given two sets A and B , their **Cartesian product** is the set

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

of **ordered pairs** of elements from A and B .

More generally, for a finite collection of sets $A_1, \dots, A_n$, their n -fold Cartesian product is the set

$$\prod_{i=1}^n A_i = A_1 \times \dots \times A_n = \{(a_1, \dots, a_n) : a_1 \in A_1, \dots, a_n \in A_n\}$$

of **ordered n -tuples** of elements from $A_1, \dots, A_n$.

So, the Cartesian product $A \times B$ consists of all possible ordered pairs (a, b) where the first element is in A and the second element is in B .

It's tempting, certainly when considering products of numerical sets such as $\mathbb{N} \times \mathbb{N}$, to expect the Cartesian product to consist of all possible products of elements. But this only makes sense when the elements have a well-defined, canonical product operation. When this isn't true (as is the case in many of the examples throughout this book), this approach fails.

The ordering is important: it's not the case in general that $A \times B = B \times A$, so the Cartesian product isn't commutative. In the next section we'll see that there is a strong relationship between $A \times B$ and $B \times A$, and we'll study in detail exactly what that relationship is.

The Cartesian product also isn't, strictly speaking, associative, since $(A \times B) \times C$ consists of ordered pairs of the form $((a, b), c)$ and $A \times$

$(B \times C)$ consists of pairs of the form $(a, (b, c))$, whereas $A \times B \times C$ consists of ordered triples of the form (a, b, c) . This distinction might seem pedantic, and it's one that we will often gloss over, but it's important to be aware of it.

From time to time we will find it necessary or useful to think about the number of elements a set has, and to that end we introduce the following terminology.

Definition A.8 A set A is said to be **finite** if it has a finite number of elements, and **infinite** otherwise.

The **cardinality** or **order** of a finite set A is the number of elements contained in A , and is denoted $|A|$.

In Section A.3 we will extend this definition, discussing the cardinality of infinite sets and investigating different types of infinity.

A.2 Functions

ONE OF THE MOST IMPORTANT concepts in the entirety of mathematics is that of a **function** or **map**. We first meet these as operations performed on real numbers, in particular polynomial expressions such as $x^2 + 3x + 1$, trigonometric functions like $\sin$, $\cos$ and $\tan$, or the exponential function e^x and the natural logarithm $\ln$.

This is all very well, as far as it goes, but we want to extend these ideas to work with arbitrary sets rather than just those consisting of real numbers.

To do this, we look carefully at what a function does: it's a device that takes an element of one set (which in the examples just listed has been some subset of $\mathbb{R}$) and gives in return an element of some other set (which might be the same or something else entirely) in a well-defined, deterministic and consistent way.

Definition A.9 Let A and B be sets. A **function** or **map** $f: A \rightarrow B$ is a rule that assigns an element of B to every element of A .

More precisely, for every $a \in A$ there is an element $f(a) \in B$.

The set A is called the **domain** of f and B is the **codomain** of f .

Look carefully at what this definition says, and also at what it doesn't say. In particular, we don't require that every element of the codomain B is mapped to by some element of the domain A , nor do we insist that distinct elements of the domain A are mapped to distinct elements of the codomain B . We will certainly consider these special cases in a

little while, but the realisation that not every element of the codomain need be mapped to inspires the following definition.

Definition A.10 Given a function $f: A \rightarrow B$, the **image** or **range** of f is the subset

$$\text{im}(f) = \{f(a) : a \in A\}$$

of the codomain B .

Example A.11 Let $f: \mathbb{R} \rightarrow \mathbb{R}$ such that $x \mapsto x^2$ for all $x \in \mathbb{R}$. Then the image

$$\text{im}(f) = \mathbb{R}_{\geq 0} = [0, \infty).$$

Example A.12 Consider the trigonometric functions $\sin: \mathbb{R} \rightarrow \mathbb{R}$, and $\cos: \mathbb{R} \rightarrow \mathbb{R}$. Then

$$\text{im}(\sin) = \text{im}(\cos) = [-1, 1].$$

Example A.13 Suppose that $f: \{a, b, c, d, e\} \rightarrow \{0, 1, 2\}$ such that $a, e \mapsto 0$ and $b, c, d \mapsto 1$. Then $\text{im}(f) = \{0, 1\}$.

These examples have made use of some common notation, which it's worth highlighting before we go any further. We use arrows to denote the direction of the mapping, so for a function $f: A \rightarrow B$ we are taking elements from A and mapping them to elements in B . But as the above examples show, there are two types of arrow in common usage. The arrow $\rightarrow$ acts on the level of sets, while the arrow $\mapsto$ acts on the level of individual elements. So the definition $f: \mathbb{R} \rightarrow \mathbb{R}; x \mapsto 3x+2$ says that the function f maps elements of $\mathbb{R}$ to elements of $\mathbb{R}$, by mapping a given element x to the element $3x+2$.

Given functions $f: A \rightarrow B$ and $g: B \rightarrow C$ we can chain them together by taking the output of f and feeding it into the input of g . We can represent this situation as

$$A \xrightarrow{f} B \xrightarrow{g} C$$

and notice that what we end up with is a well-defined way of turning an element of A into an element of C . Granted, we have to go through B to get there, but that's just an interim stage and ultimately what we have is a function $h: A \rightarrow C$. This process is called **composition**:

Definition A.14 Given functions $f: A \rightarrow B$ and $g: B \rightarrow C$, we can define the **composite function** $g \circ f: A \rightarrow C$ such that

$$(g \circ f)(a) = g(f(a))$$

for all $a \in A$.

The ordering of the functions is important: $g \circ f$ means we do f and *then* do g to the result, rather than the other way round. Applying

the functions in the reverse order need not give the same composite function; indeed the composite $f \circ g$ need not even be defined.

More generally, for two functions $f: A \rightarrow B$ and $g: C \rightarrow D$, we can define $g \circ f$ if and only if $\text{im}(f) \subseteq C$. That is, if the image of f is a subset of the domain of g .

The reason we write $g \circ f$ to denote f followed by g is because it makes the link with $g(f(a))$ clearer. In a little while we will see that this agrees with multiplication of matrices representing linear transformations.

In some respects it's less intuitive: we write the functions in the reverse order to how they're applied. Some books sidestep this issue by applying functions on the right, writing xf instead of $f(x)$. This is a more intuitive ordering for function composition, since $x(f \circ g)$ is the same as xfg , but the cost is that this notation is less intuitive in general. As a result, this convention tends to be rare now, but sometimes appears in older books.

Sometimes we can use an explicit definition for f and g to obtain an explicit definition for the composite $g \circ f$.

Example A.15 Let $f: \mathbb{R} \rightarrow \mathbb{R}$ such that $f(x) = x^2 + 5x + 1$ and $g: \mathbb{R} \rightarrow \mathbb{R}$ such that $g(x) = 2x + 3$.

Then we can define both composites $g \circ f$ and $f \circ g$. To obtain explicit expressions for these, we substitute the image of x under the action of one function into the definition of the other. So

$$\begin{aligned}(g \circ f)(x) &= g(f(x)) \\ &= g(x^2 + 5x + 1) \\ &= 2(x^2 + 5x + 1) + 3 \\ &= 2x^2 + 10x + 5, \\ (f \circ g)(x) &= f(g(x)) \\ &= f(2x + 3) \\ &= (2x + 3)^2 + 5(2x + 3) + 1 \\ &= 4x^2 + 22x + 25.\end{aligned}$$

As we can see, $f \circ g \neq g \circ f$.

The last line of this example quietly introduces the idea of two functions being equal to each other or not: two functions $f: A \rightarrow B$ and $g: C \rightarrow D$ are equal if and only if $A = C$, $B = D$ and $f(a) = g(a)$ for all $a \in A$.

Just as we declared two sets to be equal if they contained exactly the same elements, we regard two functions as equal if they have the same domains and codomains, and act exactly the same on all elements of

their domain.

One special function, albeit one that seems trivial when we first meet it, is the **identity** function.

Definition A.16 For a given set A , the **identity map** or **identity function** $\text{id}: A \rightarrow A$ is defined by $\text{id}(a) = a$ for all $a \in A$. To avoid ambiguity we might sometimes denote this function as id_A .

This might seem a bit pointless, but we should never automatically dismiss something that appears to do nothing.³ Throughout the rest of this book we see many examples of identity functions and operations that perform important rôles. One particular important property is shown by the following proposition.

Proposition A.17 For any function $f: A \rightarrow B$,

$$\text{id}_B \circ f = f = f \circ \text{id}_A.$$

Proof This is fairly straightforward, and follows almost immediately from the definitions. For any $a \in A$ we have

$$(\text{id}_B \circ f)(a) = \text{id}_B(f(a)) = f(a) \text{ and } (f \circ \text{id}_A)(a) = f(\text{id}_A(a)) = f(a).$$

Hence $\text{id}_B \circ f = f$ and $f \circ \text{id}_A = f$. □

Some function $f: A \rightarrow B$ can be “undone”, in the sense that for any element a we can apply f to get $f(a)$ and then reverse the process to recover a . More precisely, what we want is a function $g: B \rightarrow A$ such that $g(f(a)) = a$ for all $a \in A$.

Definition A.18 A function $f: A \rightarrow B$ is **invertible** if there exists a function $g: B \rightarrow A$ such that

$$g \circ f = \text{id}_A \quad \text{and} \quad f \circ g = \text{id}_B.$$

We call g the **inverse** of f and usually denote it f^{-1} , so that

$$f^{-1} \circ f = \text{id}_A \quad \text{and} \quad f \circ f^{-1} = \text{id}_B.$$

Example A.19 Let $\exp: \mathbb{R} \rightarrow \mathbb{R}_+$ be the exponential function mapping $x \mapsto e^x$, and let $\ln: \mathbb{R}_+ \rightarrow \mathbb{R}$ be the natural logarithm function. Each of these functions is the inverse of the other.

Exercise A.1 Show that inverse functions, if they exist, are unique. That is, for some function $f: A \rightarrow B$ with two inverses $g: B \rightarrow A$ and $h: B \rightarrow A$, it must be the case that $g = h$. Hence it is appropriate to refer to *the* inverse f^{-1} .

Show also that if $g: B \rightarrow A$ is the inverse of $f: A \rightarrow B$, then f is also the inverse of g . Hence $(f^{-1})^{-1} = f$.

Example A.20 Let $g: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ map $x \mapsto x^2$, and let $h: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ map $x \mapsto \sqrt{x}$. Here we follow the usual convention that $\sqrt{x}$ denotes the *positive* square root of x . Then g and h are the inverse of each other.

It's not always the case that a function is invertible. The functions in the above example are invertible because we chose the domain and codomain carefully to ensure that everything worked. The closely-related function $f: \mathbb{R} \rightarrow \mathbb{R}$ that maps $x \mapsto x^2$ fails to be invertible in

two different ways.

One problem becomes apparent when we consider what happens to 2 and -2 . Both of these square to 4, at which point we run into difficulty with the definition of an inverse function f^{-1} : we have to find some way of mapping 4 to both 2 and -2 in order to satisfy the condition $f^{-1} \circ f = \text{id}_{\mathbb{R}}$. That is, we need to define f^{-1} so that the composition $f^{-1} \circ f$ maps $2 \mapsto 2$ and $-2 \mapsto -2$. The only way we can do this is to somehow define f^{-1} so that it maps 4 to *both* 2 and -2 . But this isn't allowed: a function must map each individual element of its domain to a *single* element of its codomain, not to pairs or larger subsets of elements.

The other problem occurs when we look at the second condition $f \circ f^{-1} = \text{id}_{\mathbb{R}}$. In particular, look at what happens to -1 . No matter how we define f^{-1} and whatever we decide -1 should map to, when we then apply f to the result we get a non-negative number. So f fails the second invertibility condition as well.

The reason the first condition fails is that f maps both 2 and -2 to the same element in the codomain, which we said earlier is fine in general, but here is exactly the thing that stops f being invertible.

The second condition fails because -1 lies outside the image of f , so whatever we map it to via a candidate inverse function f^{-1} , it's not going to get mapped back to -1 when we apply f , because f doesn't map anything to -1 . Again, the existence of un-mapped-to elements in the codomain isn't a problem in general, but is precisely the thing that causes the second invertibility condition to fail.

This leads to the following two definitions.

Definition A.21 A function $f: A \rightarrow B$ is **injective** or **one to one** if every element of the codomain is the image of *at most* one element of the domain.

More precisely, f is injective if and only if $f(a_1) = f(a_2)$ implies $a_1 = a_2$.

Be careful not to get this condition the wrong way round. If $a_1 = a_2$ then $f(a_1)$ has to equal $f(a_2)$, otherwise the function isn't well-defined.

Definition A.22 A function $f: A \rightarrow B$ is **surjective** or **onto** if every element of the codomain is the image of *at least* one element of the domain.

More precisely, f is surjective if and only if for any $b \in B$ there exists some (not necessarily unique) element $a \in A$ such that $f(a) = b$.

In addition to their important invertibility properties, bijective functions are just what we need to extend our notion of cardinality to

infinite sets. The key observation is that a bijection defines an exact correspondence between its domain and codomain. We can use this as a way of counting the elements of a finite set: we construct a bijection between our chosen set and a subset of natural numbers of the form $\{1, 2, \dots, n\}$ for some $n \in \mathbb{N}$.

For example, taking the set $A = \{a, b, c, d\}$ we can define a bijection $f: A \rightarrow B$ where $B = \{1, 2, 3, 4\}$. In fact, we can define $4! = 24$ different such bijections, of which

$$f(a) = 1, \quad f(b) = 3, \quad f(c) = 4, \quad f(d) = 2$$

is one. If B has any fewer elements then no function $f: A \rightarrow B$ can be injective, and if it has any more elements then no function $f: A \rightarrow B$ can be surjective. So the only case in which a bijection can exist is when A and B have the same number of elements. And since we know that B has 4 elements, A must too.

O God, I could be bounded in a nutshell, and count myself a king of infinite space – were it not that I have bad dreams.

— William Shakespeare (1564–1616),
Hamlet II:2 (c.1600)

A.3 Counting and infinity

WE CAN EXTEND THIS APPROACH to infinite sets as well. Suppose that we take some infinite set A and manage to construct a bijection $f: A \rightarrow \mathbb{N}$. Then by the discussion above these two sets must have the same “number” of elements, although it’s usually unwise to treat infinity as a conventional number because it behaves in particularly counterintuitive ways. In fact, as we’ll see in a little while, there isn’t even just one infinity, but an entire hierarchy of them.

Our first counterintuitive consequence of this discussion is that there are as many natural numbers as there are even natural numbers:

Proposition A.23 *There is a bijection $f: 2\mathbb{N} \rightarrow \mathbb{N}$ defined by $f(n) = \frac{1}{2}n$.*

Proof Every even natural number is of the form $n = 2k$ for some $k \in \mathbb{N}$, so the function f is well-defined in the sense that every element of $2\mathbb{N}$ maps to an element of $\mathbb{N}$.

The function f is injective, since if $f(n) = f(m)$ we have $\frac{1}{2}n = \frac{1}{2}m$ and hence $n = m$. It is also surjective, since for any $m \in \mathbb{N}$ we can find $2m \in 2\mathbb{N}$ such that $f(2m) = \frac{1}{2}(2m) = m$. Therefore f is a bijection. $\square$

Equally surprisingly, there are as many integers as natural numbers:

Proposition A.24 *There is a bijection $f: \mathbb{Z} \rightarrow \mathbb{N}$ defined by*

$$f(n) = \begin{cases} 2n + 1 & \text{if } n \geq 0, \\ -2n & \text{if } n < 0. \end{cases}$$

Following the arrows yields the list

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{1}{3}, \frac{2}{2}, \frac{3}{1}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \frac{1}{5}, \frac{2}{4}, \frac{3}{3}, \frac{4}{2}, \frac{5}{1}, \frac{6}{1}, \frac{5}{2}, \frac{4}{3}, \frac{3}{4}, \frac{2}{5}, \frac{1}{6}, \dots$$

However, each rational number will occur infinitely many times (for example, $\frac{1}{2} = \frac{2}{4} = \frac{3}{6} = \dots$) so we must delete from this list the second and subsequent occurrences of every given rational number. This gives the list

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{1}{3}, \frac{3}{1}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \frac{1}{5}, \frac{5}{1}, \frac{6}{1}, \frac{5}{2}, \frac{4}{3}, \frac{3}{4}, \frac{2}{5}, \frac{1}{6}, \dots$$

which contains every single positive rational number exactly once. We can then construct a bijection $f: \mathbb{Q}_+ \rightarrow \mathbb{N}$ by mapping each rational number to the positive integer given by its position in the list. So, for example, $\frac{2}{5} \mapsto 16$ and so forth.

This function, while not constructed by means of a neat formula, is nevertheless well-defined and bijective, and so $|\mathbb{Q}_+| = |\mathbb{N}| = \aleph_0$ as claimed.

We can extend this to provide a bijection $g: \mathbb{Q} \rightarrow \mathbb{N}$ by using a similar trick to the one used in Proposition A.24: we define

$$g(q) = \begin{cases} 2f(q) + 1 & \text{if } q > 0, \\ 1 & \text{if } q = 0, \\ 2f(-q) & \text{if } q < 0. \end{cases}$$

This function g is a bijection, and thus $\mathbb{Q}$ is countable. $\square$

The obvious, naïve approach to constructing a bijection $f: \mathbb{Q} \rightarrow \mathbb{N}$ is to start in the top left-hand corner of the grid and proceed horizontally, row by row, accounting for all the rational numbers. The problem is that there are (countably) infinitely many rational numbers on the first row, so we never actually get as far as the second or subsequent rows. Cantor's method avoids this problem by traversing the grid in a zig-zag pattern that doesn't miss anything out: eventually we'll reach any given rational number in a way that doesn't happen if we follow the obvious row-by-row method.

What about the next set in our standard hierarchy of number systems? Are the real numbers countable? There are certainly infinitely many of them, because $\mathbb{R}$ contains $\mathbb{Q}$ as a proper subset, but it turns out we can't construct a bijection $f: \mathbb{R} \rightarrow \mathbb{N}$ no matter how carefully we arrange them.

The following result is also due to Cantor, and has become known as **Cantor's diagonal argument**.

Proposition A.27 *The set $\mathbb{R}$ of real numbers is not countable.*

The following proof uses a technique called **proof by contradiction** or *reductio ad absurdum*, which we'll discuss in more detail later. The

central idea is that we assume the opposite of what we're trying to prove, and then show that the logical consequences of that are so obviously incorrect or unthinkable that the original proposition must have been true all along.

Proof First we show that the interval $(0,1) = \{x : 0 < x < 1\}$ is uncountable.

Suppose for the moment that $(0,1)$ is countable after all. That means that there is a bijection $f: (0,1) \rightarrow \mathbb{N}$, and hence we can (in principle, if not in practice) write down an infinite list of all the real numbers in the open interval $(0,1)$. This yields a sequence $\mathbf{x} = (x_1, x_2, \dots)$ where x_i is the unique real number satisfying $f(x_i) = i \in \mathbb{N}$.

We can write every number $y \in (0,1)$ as an infinite decimal expansion of the form

$$y = 0.y_1y_2y_3\dots$$

where each digit $y_i \in \{0,1,2,3,4,5,6,7,8,9\}$. To avoid ambiguity we write any decimal expansion of the form

$$0.y_1y_2\dots y_k9999\dots$$

(where $y_k \neq 9$) as

$$0.y_1y_2\dots (y_k+1)0000\dots$$

instead, and we pad any finite-length decimal expansion to an infinite one by appending an infinite sequence of trailing zeros after the last nonzero digit.

Our sequence $\mathbf{x}$ of all real numbers from $(0,1)$ can be written out as infinite decimal expansions as follows:

$$\begin{aligned} x_1 &= 0.x_{11}x_{12}x_{13}x_{14}\dots \\ x_2 &= 0.x_{21}x_{22}x_{23}x_{24}\dots \\ x_3 &= 0.x_{31}x_{32}x_{33}x_{34}\dots \\ x_4 &= 0.x_{41}x_{42}x_{43}x_{44}\dots \\ &\vdots \end{aligned}$$

We now construct a new number x_0 by taking the *diagonal* digits of this list

$$\begin{aligned} x_1 &= 0.\boxed{x_{11}}x_{12}x_{13}x_{14}\dots \\ x_2 &= 0.x_{21}\boxed{x_{22}}x_{23}x_{24}\dots \\ x_3 &= 0.x_{31}x_{32}\boxed{x_{33}}x_{34}\dots \\ x_4 &= 0.x_{41}x_{42}x_{43}\boxed{x_{44}}\dots \\ &\vdots \end{aligned}$$

to get

$$0.x_{11}x_{22}x_{33}x_{44}\dots$$

and then adding 1 to each digit using modulo-10 arithmetic. This yields

$$x_0 = 0.(x_{11}+1)(x_{22}+1)(x_{33}+1)(x_{44}+1)\dots,$$

a perfectly valid real number that lies in the interval $(0, 1)$ but doesn't appear anywhere in the existing sequence $\mathbf{x}$. The reason for this is that for any $n \in \mathbb{N}$, the number x_n differs from x_0 at the n th digit.

Therefore the sequence $\mathbf{x}$ didn't contain all of $(0, 1)$, and in fact no such sequence can exist, so $(0, 1)$ is not countable after all.

Since $\mathbb{R}$ contains $(0, 1)$ as a proper subset, $\mathbb{R}$ isn't countable either. $\square$

This proves that $|\mathbb{R}| \neq \aleph_0$, and since $\mathbb{N} \subset \mathbb{R}$ we already know that $|\mathbb{N}| \leq |\mathbb{R}|$, so it must follow that $|\mathbb{N}| < |\mathbb{R}|$ and hence we need a new infinite cardinal. We say that $\mathbb{R}$ is **uncountable**, and has cardinality $\aleph_1$ (pronounced **aleph one**).

It is possible to construct hierarchies of successively larger infinite cardinals, but we won't go into the details. For a very readable discussion of infinity and related matters see the book *Infinity and the Mind* by the mathematician, logician, artist and science fiction writer Rudy Rucker.⁴ The Argentinian magic realist writer Jorge Luis Borges (1899–1986) explores ideas of infinity in several of his stories, in particular *The Library of Babel*,⁵ *The Aleph*⁶ and *The Book of Sand*.⁷

Bijjective functions also enable us to properly answer the question about the commutativity and associativity of cartesian products. As noted earlier, in the discussion following Definition A.7, in general $A \times B \neq B \times A$ for arbitrary sets A and B . There is, however, an obvious bijection $f: A \times B \rightarrow B \times A$ given by $f(a, b) = (b, a)$ for all $a \in A$ and $b \in B$.

Similarly, although $A \times (B \times C) \neq (A \times B) \times C$, we can define a bijection $g: A \times (B \times C) \rightarrow (A \times B) \times C$ by $g((a, b), c) = (a, (b, c))$ and therefore it makes sense in almost all cases to regard these as effectively the same as each other and the threefold Cartesian product $A \times B \times C$.

⁴R Rucker, *Infinity and the Mind*, Princeton University Press (2005).

⁵JL Borges, *The Library of Babel*, in: *Labyrinths*, Penguin (1970) 78–86.

⁶JL Borges, *The Aleph*, in: *The Aleph and Other Stories*, Penguin (2000) 118–133.

⁷JL Borges, *The Book of Sand*, in: *The Book of Sand and Shakespeare's Memory*, Penguin (2001) 89–93.

A.4 Relations

EARLIER, WE NOTED THAT a set has, by default, no particular predetermined structure. Most of the rest of this book is concerned with the study of some particularly important structures that arise naturally in various contexts. Now we will examine another class of structures that can be defined on sets.

We often find it necessary or useful to compare individual elements

of sets in various ways. For example, sometimes we need to know in what circumstances two arbitrary elements are equal to each other. Or we might be interested in whether one element is greater than another (if the set in question actually has a coherent concept of “greater than” or “less than”).

More generally, we sometimes find it necessary to study other connections between elements. In number theory, for example, we are often interested in whether some integer is a factor of another.

We’d like to formalise this idea and see what important properties arise. To start with, any two elements of a set are either related in a specific way or they aren’t. So what we want is a way of assigning a value of either “true” or “false” to a given ordered pair of elements, depending on whether those elements are related in the specified way or not. One way of looking at this is as a function $f: A \times A \rightarrow \{\text{true}, \text{false}\}$.

Alternatively, we can define a relation by listing the ordered pairs of related elements. This approach views a relation as a subset of the cartesian product $A \times A$. Both approaches are valid and equivalent, and each has its own advantages, so we present both.

Definition A.28 A **(binary) relation** on a set A is a function $f: A \times A \rightarrow \{\text{true}, \text{false}\}$, and we say that two given elements $a, b \in A$ are **related** if $f(a, b) = \text{true}$.

Equivalently, a **(binary) relation** on a set A is a subset $B \subseteq A \times A$, and we say that two given elements $a, b \in A$ are **related** if $(a, b) \in B$.

The notation $f(a, b) = \text{true}$ or $(a, b) \in B$ is rather cumbersome, and obscures the comparative aspect of the relation under discussion, so in most cases we will represent a given relation with a symbol such as $\sim$ placed between the two elements. This enables us to write more intuitive expressions such as $a \sim b$ if a is related to b , and $a \not\sim b$ if not.

Example A.29 For any set A we can define the **equality** relation = defined by

$$f(a, b) = \begin{cases} \text{true} & \text{if } a = b, \\ \text{false} & \text{if } a \neq b. \end{cases}$$

Example A.30 Let $f: \mathbb{N} \times \mathbb{N} \rightarrow \{\text{true}, \text{false}\}$ be defined by

$$f(m, n) = \begin{cases} \text{true} & \text{if } m \text{ is a factor of } n, \\ \text{false} & \text{otherwise.} \end{cases}$$

This determines the relation $|$ or “divides”: we say that $m|n$ if and only if $f(m, n) = \text{true}$.

Another important relation is the notion of an ordering on a set:

Example A.31 Suppose $A \subseteq \mathbb{R}$. Then the “less than” relation $<$ is given by

$$f(a, b) = \begin{cases} \text{true} & \text{if } a < b, \\ \text{false} & \text{if } a \geq b. \end{cases}$$

We can define the “less than or equal” relation $\leq$ in a similar way.

Each of these relations have subtly different properties. Is it always the case that a given element is related to itself? In the case of $=$ it so happens that $a = a$ for any element $a \in A$, but it’s not the case that $a < a$ in general (or indeed at all) for subsets $A \subseteq \mathbb{R}$, although clearly $a \leq a$ in all cases. Furthermore, $n|n$ for any $n \in \mathbb{N}$.

Does the order of the elements matter? If $a \sim b$ then is it necessarily true that $b \sim a$? If $a = b$ then clearly $b = a$ in all cases. But if $a < b$ it’s certainly not the case that $b < a$, and if $m|n$ it’s not true that $n|m$ (except when $m = n$).

What about comparisons of three arbitrary elements? If $a = b$ and $b = c$ then clearly $a = c$. Also, if $a < b$ and $b < c$ then it’s necessarily the case that $a < c$. And if $p|q$ and $q|r$ then $p|r$.

Are there any relations that don’t satisfy this third property? Probably the best known example is the old game of “**rock–paper–scissors**” (sometimes called **roshambo** or **ick ack ock**).⁸ We can define a relation “beats” on the set $\{\text{rock, paper, scissors}\}$ and then observe that although rock beats scissors, and scissors beat paper, rock doesn’t beat paper. This relation also doesn’t satisfy the other two conditions.

⁸ Here, rock beats scissors (because the rock blunts the scissors), scissors beat paper (because the scissors cut the paper) and paper beats rock (because the paper wraps the rock). Every other outcome is a draw.

This game can be generalised to any odd number of players: there is a popular five-player variant called “rock–paper–scissors–lizard–Spock”.

Definition A.32 A relation $\sim$ defined on some set A is said to be

- (i) **reflexive** if $a \sim a$ for all $a \in A$,
- (ii) **symmetric** if $a \sim b$ implies that $b \sim a$ for all $a, b \in A$, and
- (iii) **transitive** if $a \sim b$ and $b \sim c$ together imply that $a \sim c$ for all $a, b, c \in A$.

So, in this terminology, $=$ is reflexive, symmetric and transitive; $<$ is transitive only; and $\leq$ and $|$ are both reflexive and transitive but not symmetric. The relation “beats” in rock–paper–scissors is neither reflexive, symmetric nor transitive.

A relation which satisfies all three of these conditions is in some sense stronger or more exclusive than one that satisfies only some or none of them. We can view such a relation as a kind of generalised notion of equality or equivalence:

Definition A.33 A relation $\sim$ on a set A is said to be an **equivalence relation** if it is symmetric, reflexive and transitive.

The motivating example of an equivalence relation is, of course, the

equality relation $=$, but there are others that arise naturally in certain contexts. In Section 2.2 we formulate a particularly important class of equivalence relations on groups, and use it to prove Lagrange's Theorem.⁹ Probably the best-known example of an equivalence relation, apart from $=$ itself, is that of congruence modulo n .

⁹ Theorem 2.30, page 54.

Definition A.34 Two integers $a, b \in \mathbb{Z}$ are **congruent modulo n** for some positive integer $n \in \mathbb{N}$ if $(a-b)|n$. Or, equivalently, if they both have the same remainder when divided by n .

We write $a \equiv_n b$ or $a \equiv b \pmod{n}$.

This is an equivalence relation on the set of integers.

A.5 Number theory

MUCH OF ABSTRACT ALGEBRA is concerned with extending and generalising facts about the integers to wider classes of objects, such as polynomials, symmetry operations, matrices and so forth. In this section we review a few elementary number-theoretic results used elsewhere in the book. First we introduce a couple of basic definitions.

Definition A.35 Let $a, b \in \mathbb{Z}$. Then a is a **factor** or **divisor** of, or **divides** b , if $b = ka$ for some integer $k \in \mathbb{Z}$. We write $a|b$ if this is so, and $a \nmid b$ if not.

The **greatest common divisor** or **highest common factor** of two positive integers $m, n \in \mathbb{N}$ is the largest positive integer $d \in \mathbb{N}$ that divides both m and n . We write this as $\gcd(m, n)$. If $\gcd(m, n) = 1$ then m and n are said to be **coprime** or **relatively prime**.

The ancient Greek mathematician Euclid of Alexandria (fl. 300BC) gives an algorithm for finding the greatest common divisor of two positive integers. It appears as Propositions 1 and 2 in Book VII (and later, in a slightly different form, as Propositions 2 and 3 in Book X) of his celebrated treatise $\Sigma\tau\omicron\iota\chi\epsilon\iota\alpha$ (*Elements*), although it almost certainly dates back even earlier than that.

In order to prove the validity of this algorithm, we need a basic fact about natural numbers, namely that the process of division with remainders, which most of us learn at primary school, actually works.¹⁰

Theorem A.37 (The Division Theorem) *Let $a, b \in \mathbb{N}$. Then there exist unique integers $q, r \in \mathbb{Z}$ such that $a = bq + r$ and $0 \leq r < b$.*

Proof First, we prove the existence of q and r . If $a = b$ then obviously $a = 1b + 0$, so $q = 1$ and $r = 0$ suffice.

God created the integers; all else is the work of man.

— Leopold Kronecker (1823–1891),
lecture at Göttingen (1886),
quoted by Heinrich Weber (1842–1913),
Jahresbericht der Deutschen
Mathematiker-Vereinigung 2 (1892) 19

¹⁰ This proof makes use of the method of **Proof by Induction**, which relies on (indeed, its validity is formally equivalent to) the **Well-Ordering Principle**:

Axiom A.36 (The Well-Ordering Principle) *Let $S \subseteq \mathbb{N}$ such that $S \neq \emptyset$. That is, let S be a nonempty set of natural numbers. Then S has a least element.*

Depending on exactly which version of axiomatic set theory one is working with, this statement is either a basic axiom or a provable proposition. We will cheerfully ignore such concerns in the rest of this book.

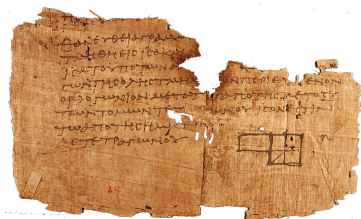


Wikimedia Commons / Raphael (1483–1520), detail from *The School of Athens* (1509–1510)

Almost nothing is known of the life of the ancient Greek mathematician **Euclid of Alexandria** (fl. 300BC), the few historical references having been written centuries later. He is believed to have lived during the reign of Ptolemy I of Egypt (c.367–c.283BC), and there are six surviving works that are credibly attributed to him, the best-known of which is the *Elements* (Στοιχεῖα).

This work, consisting of thirteen books of definitions, axioms and propositions on plane geometry, spatial geometry and number theory, is perhaps one of the most influential mathematical texts of all time: for several centuries it was part of the standard Western university and school curriculum.

One of the earliest surviving fragments is contained in the Oxyrhynchus Papyrus, dated to around 100AD, but probably the earliest surviving complete texts are an edition prepared in the fourth century AD by Theon of Alexandria (c.335–c.405AD) and a Byzantine palimpsest dating from about 900AD discovered in 1808 by the French historian François Peyrard (c.1760–1822).



Wikimedia Commons / Oxyrhynchus Papyrus, fragment 29

If $a \neq b$ then we proceed by induction on a . For $a=1$ (and thereby $b>1$) we have

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

with $q = 0$ and $r = 1 < b$.

Now suppose that $a = qb + r$ for some $q, r \in \mathbb{Z}$ with $0 \leq r < b$. Then adding 1 to both sides of this equation we obtain

$$a + 1 = qb + r + 1.$$

Since $r < b$ it follows that $r + 1 \leq b$. If $r + 1 = b$ then

$$a + 1 = (q + 1)b + 0,$$

and if $r + 1 < b$ we have

$$a + 1 = qb + (r + 1).$$

In each case $a+1$ can be expressed in the desired way, and hence it follows by induction that any $a \in \mathbb{N}$ can be decomposed in this way too.

To show uniqueness, suppose that there exist two other integers $s, t \in \mathbb{Z}$ such that $a = sb + t$, with $0 \leq t < b$. Then $sb + t = qb + r$, which implies that $(q - s)b = t - r$.

This means that $(t - r)$ is divisible by b . Since $0 \leq r, t < b$ it follows that $0 \leq |t - r| < b$, which means that

$$-(b - 1) \leq t - r \leq b - 1.$$

The only integer in this range divisible by b is 0, so $t - r = 0$ and hence $t = r$.

Now we have $qb + r = sb + t = sb + r$, and hence $qb = sb$, from which we find that $q = s$. Hence q and r are unique. $\square$

We can now state Euclid's Algorithm, and prove that it works.

Theorem A.38 (Euclid's Algorithm) *Let $a, b \in \mathbb{N}$. Then there exist unique integers $q_1, \dots, q_{n+1}$ (the **quotients**) and $b > r_1 > \dots > r_{n+1} = 0$ (the **remainders**) such that*

$$a = q_1b + r_1,$$

$$b = q_2r_1 + r_2,$$

$$r_1 = q_3r_2 + r_3,$$

$$\vdots$$

$$r_{n-2} = q_nr_{n-1} + r_n,$$

$$r_{n-1} = q_{n+1}r_n + r_{n+1},$$

and $\gcd(a, b) = r_n$, the last nonzero remainder.

Proof First we apply the Division Theorem (Theorem A.37) to the pair a, b to find the unique integers q_1, r_1 satisfying $a = q_1b + r_1$ and $0 \leq r_1 < b$. If $r_1 = 0$ then we stop. Otherwise, we apply the Division Theorem again to the pair b, r_1 to find the unique integers q_2, r_2 such that $b = q_2r_1 + r_2$. If $r_2 = 0$ then we stop, otherwise we apply the Division Theorem yet again to the pair r_1, r_2 .

Continuing this process, we obtain a strictly decreasing sequence of remainders $r_1 > r_2 > \cdots$ which are all less than b and strictly non-negative. Hence, after finitely many steps, we reach $r_{n+1} = 0$.

The last step is therefore

$$r_{n-1} = q_{n+1}r_n + 0$$

and so $r_{n-1} = q_{n+1}r_n$, which means that $r_n | r_{n-1}$.

The penultimate step, similarly, is

$$r_{n-2} = q_nr_{n-1} + r_n = q_nq_{n+1}r_n + r_n = (q_nq_{n+1} + 1)r_n$$

and so $r_n | r_{n-2}$. Continuing backwards, we find that r_n divides all of the other remainders $r_{n-1}, \dots, r_1$. Furthermore, since $r_n | r_2$ and $r_n | r_1$, it follows that r_n also divides $b = q_2r_1 + r_2$. And since $r_n | b$, it also follows that r_n divides $a = q_1b + r_1$. Thus r_n is a factor of both a and b . All that remains is to show that r_n is the largest such integer; that is, $r_n = \gcd(a, b)$. Suppose $d \in \mathbb{N}$ is another common factor of a and b . Rearranging the equations obtained via the Division Theorem, we get

$$\begin{aligned} r_1 &= a - q_1b, \\ r_2 &= b - q_2r_1, \\ r_3 &= r_1 - q_3r_2, \\ &\vdots \\ r_n &= r_{n-2} - q_nr_{n-1}. \end{aligned}$$

Now suppose $d | a$ and $d | b$. Then from the first equation we find that $d | r_1$, from the second that $d | r_2$, from the third that $d | r_3$, and so on. Finally, we find that $d | r_{n-2}$ and $d | r_{n-1}$, so $d | r_n$.

Therefore $d \leq r_n$, and hence r_n is the highest such factor. That is, $r_n = \gcd(a, b)$ as claimed. $\square$

Corollary A.39 Let $a, b \in \mathbb{N}$, and let $d = \gcd(a, b)$. Then there exist integers $s, t \in \mathbb{Z}$ such that $d = sa + tb$.

Proof From the latter part of the above proof, we know that Euclid's Algorithm ensures the existence of integers $q_1, \dots, q_{n+1}$ and $b > r_1 > \cdots > r_{n+1} > 0$ such that

$$\begin{aligned} r_1 &= a - q_1b, \\ r_2 &= b - q_2r_1, \end{aligned}$$

$$\begin{aligned}
r_3 &= r_1 - q_3 r_2, \\
&\vdots \\
r_{n-1} &= r_{n-3} - q_{n-1} r_{n-2}, \\
d = r_n &= r_{n-2} - q_n r_{n-1}.
\end{aligned}$$

Substituting the penultimate expression into the last one, we obtain

$$\begin{aligned}
d = r_n &= r_{n-2} - q_n(r_{n-3} - q_{n-1}r_{n-2}) \\
&= (1 + q_n q_{n-1})r_{n-2} - q_n r_{n-3}.
\end{aligned}$$

Hence $d = r_n$ can be expressed as a $\mathbb{Z}$ -linear combination of r_{n-3} and r_{n-2} . Now we use the antepenultimate expression on the list to substitute for r_{n-2} and thereby express d as a $\mathbb{Z}$ -linear combination of r_{n-4} and r_{n-3} .

Continuing this process, we eventually obtain an expression for d as a $\mathbb{Z}$ -linear combination of a and b , as required. $\square$

The following important fact about coprime integers is easy to prove, and will be used elsewhere in the book.

Proposition A.40 *Let a and b be coprime integers. Then if a divides bq then a must divide q .*

Proof If a and b are relatively prime, then by Corollary A.39 there exist integers s and t such that

$$sa + tb = 1.$$

Multiplying both sides of this equation by q we get

$$saq + tbq = q.$$

Since a obviously divides saq , and by hypothesis a also divides bq , the entire left hand side of this equation is divisible by a . Hence a must also divide the right hand side as well, which means that a divides q as claimed. $\square$

A.6 Real analysis

Theorem A.41 (The Intermediate Value Theorem) *Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be continuous on the closed interval $[a, b]$ for some $a, b \in \mathbb{R}$. Then if $f(a)$ and $f(b)$ have opposite signs, there exists some $c \in (a, b)$ for which $f(c) = 0$.*

Theorem A.42 (Rolle's Theorem) *Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b) , and suppose that $f(a) = f(b)$. Then there exists some $c \in (a, b)$ for which $f'(c) = 0$.*

A.7 *Linear algebra*

Index

- Abel, Niels Henrik, 7
Araucaria, *see* Graham, John
The Art of War, 297
Āryabhaṭa, 297
Āryabhaṭīya, 297
- Bernoulli, Johann, 11
binary operation, 2
Brahmagupta, 5
- Cardano, Girolamo, 5
commutative, 3
- Darboux, Jean Gaston, 20
Diophantus, 5
- Engel, Friedrich, 20
Euler, Leonhard, 11
- totient function, 15
- Feynman, Richard, 1
Fibonacci, 5, 6, 297
sequence, 6
- Galois, Évariste, 7
Graham, John, 27
- Jiuzhang suanshu*, *see* *Nine Chapters on the Mathematical Art*
Jordan, Camille, 20
- Klein, Felix, 17, 20
- Leonardo of Pisa, *see* Fibonacci
Liber Abaci, 6, 297
Lie, Marius Sophus, 20
- magma, 3
Mahāvira, 5
Mathematical Classic of Sun Tzu, 297
Maugham, W Somerset, 303
- Nine Chapters on the Mathematical Art*, 5
- Pauli, Wolfgang, 21
Plücker, Julius, 17
Poincaré, Henri, 22
- Sun Tzu, 297
Sun Tzu Suan Ching, *see* *Mathematical Classic of Sun Tzu*
- Young, Grace Chisholm, 17

