Frequently Used Notation

$\lvert A \rvert$	the order of the set A
$\lvert x \rvert$	the order of the element x
$\mathbb{Z}$, $\mathbb{Z}^+$	the integers, the positive integers
$\mathbb{Q}$, $\mathbb{Q}^+$	the rational numbers, the positive rational numbers
$\mathbb{R}$, $\mathbb{R}^+$	the real numbers, the positive real numbers
$\mathbb{C}$	the complex numbers
$\mathbb{Z}/n\mathbb{Z}$	the integers modulo n
$(\mathbb{Z}/n\mathbb{Z})^\times$	the (multiplicative group of) invertible integers modulo n
$A \times B$	the direct or Cartesian product of A and B
$A \oplus B$	the direct sum of A and B
X^r	the direct or Cartesian product of r copies of X
$f^{-1}(A)$	the inverse image or preimage of A under f
(a, b)	the greatest common divisor of a, b also the ideal generated by a, b
$a \mid b$	a divides b
$H \leq G$	H is a subgroup of G
$H < G$	H is a proper subgroup of G
Z_n	the cyclic group of order n
D_{2n}	the dihedral group of order $2n$
S_Ω	the symmetric group on the set Ω
S_n	the symmetric group on n letters
A_n	the alternating group on n letters
Q_8	the quaternion group of order 8
V_4	the Klein 4-group
QD_{16}	the quasidihedral group of order 16
$\mathbb{F}_N$	the finite field of N elements
$GL_n(F)$	the general linear group
$SL_n(F)$	the special linear group
$\mathrm{Aut}(G)$	the automorphism group of G
$C_G(A)$	the centralizer in G of A
$N_G(A)$	the normalizer in G of A
$Z(G)$	the center of G
G_s	the stabilizer in G of s
x^g, H^g	the (right) conjugate of x (respectively, H) by g

$\langle A \rangle$	the group generated by the set A
$\langle x \rangle$	the group generated by the element x
$G = \langle . \ . \ . \mid . \ . \ . \rangle$	generators and relations (presentation) for G
$\ker \varphi$	the kernel of φ
$\operatorname{im} \varphi$	the image of φ
$N \trianglelefteq G$	N is normal in G
$\lvert G : H \rvert$	the index of H in G
H char G	H is characteristic in G
$Syl_p(G)$	the Sylow p-subgroups of G
n_p	the number of Sylow p-subgroups of G
$[x, y]$	the commutator of x, y
$H \rtimes K$	the semidirect product of H and K
$F[x]$	polynomials in x with coefficients in F
$M_n(R)$	$n \times n$ matrices over R
$M_{n \times m}(R)$	$n \times m$ matrices over R
$M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$	the matrix of the linear transformation φ with respect to bases $\mathcal{B}, \mathcal{E}$
$m_T(x)$	the minimal polynomial of the linear transformation T
$c_T(x)$	the characteristic polynomial of the linear transformation T
$\operatorname{ch}(F)$	the characteristic of F
K/F	the field K is an extension of the field F
$[K : F]$	the degree of the field extension K/F
$F(\alpha)$, $F(\alpha, \beta)$, etc.	the field generated over F by α or α, β, etc.
$m_{\alpha,F}(x)$	the minimal polynomial of α over F
$\operatorname{Aut}(K)$	the group of automorphisms of a field K
$\operatorname{Aut}(K/F)$	the group of automorphisms of a field K fixing the field F
$\operatorname{Gal}(K/F)$	the Galois group of the extension K/F
$\mathrm{N}_{K/F}(\alpha)$	the norm of α from K to F
$\mathrm{Tr}_{K/F}(\alpha)$	the trace of α from K to F
$tr(A)$	the trace of the matrix A
FG	the group ring of G over F
$GL(V)$	the group of invertible linear transformations of V
(θ, ψ)	the inner product of the characters θ, ψ
$\lVert \theta \rVert$	the norm of the character θ
$\operatorname{Ind}_H^G(\psi)$	character of the induced representation

Dummit, David Steven.
Abstract algebra / David S. Dummit, Richard M. Foote.
p. cm.
ISBN 0-13-004771-6
1. Algebra, Abstract. I. Foote, Richard M., 1950-
II. Title.
QA162.D85 1991
512'.02—dc20 90-35456
CIP

Editorial Production/Supervision: **Ed Thomas**
Cover design: **George Cornell**
Manufacturing buyer: **Paula Massenaro**

© 1991 by Prentice-Hall, Inc.
A Division of Simon & Schuster
Englewood Cliffs, New Jersey 07632

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-004771-6

PRENTICE-HALL INTERNATIONAL (UK) LIMITED, *London*
PRENTICE-HALL OF AUSTRALIA PTY. LIMITED, *Sydney*
PRENTICE-HALL CANADA INC., *Toronto*
PRENTICE-HALL HISPANOAMERICANA, S.A., *Mexico*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*
SIMON & SCHUSTER ASIA PTE. LTD., *Singapore*
EDITORA PRENTICE-HALL DO BRASIL, LTDA., *Rio de Janeiro*

UNIVERSITY OF
LIBRARY

Contents

Preface *xii*

Preliminaries *1*

0.1 Basics 1

0.2 Properties of the Integers 4

0.3 $\mathbb{Z}/n\mathbb{Z}$: The Integers Modulo n 8

Group Theory **13**

Chapter 1 *Introduction to Groups* *17*

1.1 Basic Axioms and Examples 17

1.2 Dihedral Groups 24

1.3 Symmetric Groups 28

1.4 Matrix Groups 33

1.5 The Quaternion Group 34

1.6 Homomorphisms and Isomorphisms 35

1.7 Group Actions 40

Chapter 2 ***Subgroups*** ***45***

2.1 Definition and Examples 45

2.2 Centralizers and Normalizers, Stabilizers and Kernels 48

2.3 Cyclic Groups and Cyclic Subgroups 53

2.4 Subgroups Generated by Subsets of a Group 60

2.5 The Lattice of Subgroups of a Group 65

Chapter 3 ***Quotient Groups and Homomorphisms*** ***73***

3.1 Definitions and Examples 73

3.2 More on Cosets and Lagrange's Theorem 89

3.3 The Isomorphism Theorems 96

3.4 Composition Series and the Hölder Program 102

3.5 Transpositions and the Alternating Group 107

Chapter 4 ***Group Actions*** ***113***

4.1 Group Actions and Permutation Representations 113

4.2 Groups Acting on Themselves by Left Multiplication—Cayley's Theorem 119

4.3 Groups Acting on Themselves by Conjugation—The Class Equation 123

4.4 Automorphisms 134

4.5 Sylow's Theorem 140

4.6 The Simplicity of A_n 151

Chapter 5 ***Direct and Semidirect Products and Abelian Groups*** ***153***

5.1 Direct Products 153

5.2 The Fundamental Theorem of Finitely Generated Abelian Groups 159

5.3 Table of Groups of Small Order 168

5.4 Recognizing Direct Products 170

5.5 Semidirect Products 176

Chapter 6 Further Topics in Group Theory ***190***

6.1 p-Groups, Nilpotent Groups and Solvable Groups 190

6.2 Applications in Groups of Medium Order 203

6.3 A Word on Free Groups 216

Ring Theory **224**

Chapter 7 Introduction to Rings ***225***

7.1 Basic Definitions and Examples 225

7.2 Examples: Polynomial Rings, Matrix Rings, and Group Rings 233

7.3 Ring Homomorphisms and Quotient Rings 239

7.4 Properties of Ideals 250

7.5 Rings of Fractions 261

7.6 The Chinese Remainder Theorem 267

Chapter 8 Euclidean Domains, Principal Ideal Domains and Unique Factorization Domains ***271***

8.1 Euclidean Domains 271

8.2 Principal Ideal Domains (P.I.D.s) 279

8.3 Unique Factorization Domains (U.F.D.s) 282

Chapter 9 Polynomial Rings ***291***

9.1 Definitions and Basic Properties 291

9.2 Polynomial Rings Over Fields I 294

9.3 Polynomial Rings that are Unique Factorization Domains 297

9.5 Polynomial Rings Over Fields II 307

Modules and Vector Spaces **314**

Chapter 10 Introduction to Module Theory ***315***

10.1 Basic Definitions and Examples 315

10.2 Quotient Modules and Module Homomorphisms 322

10.3 Generation of Modules, Direct Sums, and Free Modules 327

Chapter 11 Vector Spaces ***335***

11.1 Definitions and Basic Theory 335

11.2 The Matrix of a Linear Transformation 341

11.3 Dual Vector Spaces 357

11.4 Determinants 360

Chapter 12 Modules Over Principal Ideal Domains ***367***

12.1 The Basic Theory 369

12.2 The Rational Canonical Form 383

12.3 The Jordan Canonical Form 402

Field Theory and Galois Theory **421**

Chapter 13 Field Theory ***422***

13.1 Basic Theory of Field Extensions 422

13.2 Algebraic Extensions 432

13.3 Classical Straightedge and Compass Constructions 443

13.4 Splitting Fields and Algebraic Closures 448

13.5 Separable and Inseparable Extensions 458

464

Chapter 14 Galois Theory *471*

14.1 Basic Definitions 471

14.2 The Fundamental Theorem of Galois Theory 481

14.3 Finite Fields 499

14.4 Composite Extensions and Simple Extensions 505

14.5 Cyclotomic Extensions and Abelian Extensions over $\mathbb{Q}$ 510

14.6 Galois Groups of Polynomials 520

14.7 Solvable and Radical Extensions: Insolvability of the Quintic 538

14.8 Computation of Galois Groups over $\mathbb{Q}$ 553

14.9 Transcendental Extensions, Inseparable Extensions, Infinite Galois Groups 558

Introduction to the Representation Theory of Finite Groups **569**

Chapter 15 Representation Theory and Character Theory *571*

15.1 Linear Actions and Modules Over Group Rings 571

15.2 Projective and Injective Modules 585

15.3 Statement of Wedderburn's Theorem and Some Consequences 590

15.4 Character Theory and the Orthogonality Relations 600

Chapter 16 Examples and Applications of Character Theory *615*

16.1 Characters of Groups of Small Order 615

16.2 Theorems of Burnside and Hall 621

16.3 An Introduction to the Theory of Induced Characters 628

Appendix: ***Cartesian Products and Zorn's Lemma*** ***643***

Index ***650***

Preface

This book evolved out of notes by the authors from courses given at various universities over a period of about thirteen years. The backgrounds of students in these courses were quite diverse, ranging from freshman and sophomore undergraduates to beginning graduate students, and included computer science and other science and engineering majors in addition to mathematics majors. The expectations of the students were quite varied. For some, the algebra course was to be an introduction with no follow up (some only took one semester of the full year sequence) and for some the course was intended to be the foundation for future work in mathematics. We felt that all of these students should be given some insight into the main themes in abstract algebra and some understanding of how these themes provide a unifying framework for the study of the basic algebraic structures: groups, rings, and fields. One of our guiding principles has therefore been to introduce as early as possible notions such as homomorphisms, isomorphisms, actions, classifications, and the notion of the structure of an algebraic object. We point out how these themes recur as we pass from the study of one algebraic system to another and provide a natural flow for the development of basic abstract algebra. To a certain extent this has influenced our choice of material as well. We have made no attempt at being encyclopedic, but rather we have delved more deeply into a few areas in order that students may achieve a better grasp of how the structure of algebraic objects is unravelled in greater depth. Wherever possible we have attempted to broaden students' horizons by pointing out connections of one subject to other areas of algebra and mathematics and by mentioning results at the forefront of research. We also introduce new material in a form which students do not subsequently have to dislodge in favor of more

sophisticated versions. So, for example, permutations are introduced in cycle notation format and quotients are developed as fibers of homomorphisms. In this way students are immediately exposed to concepts the way an algebraist (or mathematician) envisions and works with them.

The book contains a wealth of examples and exercises. The exercises range from routine computations to fairly sophisticated theoretical ones. All the exercises, however, are within the scope of the text and hints are given [in brackets] where we felt they were needed to make them accessible. The main results in the text do not rely on material in the exercises, although some details in proofs in the text are occasionally left to the reader and some examples use results from the exercises. In many instances, new material is introduced first in the exercises (often a few sections before it appears in the text) so that students may obtain an easier introduction to it by doing these exercises (e.g. Lagrange's Theorem appears in the exercises in Section 1.7 and in the text in Section 3.2). The exercises also contain a considerable amount of material not covered in the text. Some exercises lead through a series of steps to new results such as Wedderburn's Theorem on finite division rings or Hilbert's Basis Theorem. There sometimes appear sequences of exercises on one topic; these problems are prefaced with some remarks and are structured so that they may be read as text without actually doing the exercises. Some of these sequences are reviews (such as the exercises on Gauss–Jordan elimination in Chapter 11) and others contain new material relating to another area (such as the exercises on the exponential of a matrix and its application to systems of differential equations in Section 12.3). In addition, a number of exercises describe useful computational techniques (for example, Berlekamp's Algorithm for factoring polynomials mod p in Section 14.3).

Notationally, Proposition 2 in Chapter 4 is referred to simply as Proposition 2 within Chapter 4 and as Proposition 4.2 in other chapters.

It is not usually possible to cover the entire text in one academic year, although the authors have been able to cover the bulk of the material in one year, including all of Part III and much of Part V. A basic introductory (one year) course should include Chapters 1 to 3, Sections 4.1 to 4.3, Sections 5.1 to 5.3, Chapters 7 to 9, Sections 10.1, 11.1 and Chapters 13 and 14, returning to pick up Sections 4.4 to 4.6 if they are not covered initially. Parts III and V may also be used as a well integrated one semester course. We note that linear algebra is not assumed to be a prerequisite although knowledge of it does increase the repertoire of examples in group theory. The basic linear algebra covered in Chapter 11 provides the background mainly for the theory of canonical forms in Chapter 12.

Finally, we would like to thank D. Dorman, H. Kisilevsky, J. Sands, L. Simons and our students over the past few years for helpful suggestions and comments.

D. Dummit
R. Foote

Preliminaries

Some results and notation that are used throughout the text are collected in this chapter for convenience. Students may wish to review this chapter quickly at first and then read each section more carefully again as the concepts appear in the course of the text.

0.1 BASICS

The basics of set theory: sets, $\cap, \cup, \in$, etc. should be familiar to the reader. Our notation for subsets of a given set A will be

$$B = \{a \in A \mid \ldots \text{ (conditions on a) } \ldots\}.$$

The *order* or *cardinality* of a set A will be denoted by $|A|$. If A is a finite set the order of A is simply the number of elements of A.

It is important to understand how to test whether a particular $x \in A$ lies in a subset B of A (cf. Exercises 1 to 4). The *Cartesian product* of two sets A and B is the collection $A \times B = \{(a, b) \mid a \in A, b \in B\}$, of ordered pairs of elements from A and B.

We shall use the following notation for some common sets of numbers:

(1) $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$ denotes the *integers* (the $\mathbb{Z}$ is for the German word for numbers: "Zahlen").

(2) $\mathbb{Q} = \{a/b \mid a, b \in \mathbb{Z}, b \neq 0\}$ denotes the *rational numbers* (or *rationals*).

(3) $\mathbb{R} = \{$ all decimal expansions $\pm\, d_1 d_2 \ldots d_n . a_1 a_2 a_3 \ldots \}$ denotes the *real numbers* (or *reals*).

(4) $\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}, i^2 = -1\}$ denotes the *complex numbers*.

(5) $\mathbb{Z}^+, \mathbb{Q}^+$ and $\mathbb{R}^+$ will denote the positive (nonzero) elements in $\mathbb{Z}, \mathbb{Q}$ and $\mathbb{R}$, respectively.

We shall use the notation $f : A \to B$ or $A \xrightarrow{f} B$ to denote a function f from A to B and the value of f at a is denoted $f(a)$ (i.e. we shall apply all our functions on the left). We use the words *function* and *map* interchangeably. The set A is called the *domain* of f and B is called the *codomain* of f. The notation $f : a \mapsto b$ or $a \mapsto b$ if f is understood indicates that $f(a) = b$, i.e. the function is being specified on *elements*.

If the function f is not specified on elements it is important in general to check that f is *well defined*, i.e. is unambiguously determined. For example, if the set A is the union of two subsets A_1 and A_2 then one can try to specify a function from A to the set $\{0, 1\}$ by declaring that f is to map everything in A_1 to 0 and is to map everything in A_2 to 1. This unambiguously defines f unless A_1 and A_2 have elements in common (in which case it is not clear whether these elements should map to 0 or to 1). Checking that this f is well defined therefore amounts to checking that A_1 and A_2 have no intersection.

The set

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

is a subset of B, called the *range* or *image* of f (or the *image of A under f*). For each subset C of B the set

$$f^{-1}(C) = \{a \in A \mid f(a) \in C\}$$

consisting of the elements of A mapping into C under f is called the *preimage* or *inverse image* of C under f. For each $b \in B$, the preimage of $\{b\}$ under f is called the *fiber* of f over b. Note that f^{-1} is not in general a function and that the fibers of f generally contain many elements since there may be many elements of A mapping to the element b.

If $f : A \to B$ and $g : B \to C$, then the composite map $g \circ f : A \to C$ is defined by

$$(g \circ f)(a) = g(f(a)).$$

Let $f : A \to B$.

(1) f is *injective* or is an *injection* if whenever $a_1 \neq a_2$, then $f(a_1) \neq f(a_2)$.

(2) f is *surjective* or is a *surjection* if for all $b \in B$ there is some $a \in A$ such that $f(a) = b$, i.e. the image of f is *all* of B. Note that since a function always maps onto its range (by definition) it is necessary to specify the codomain B in order for the question of surjectivity to be meaningful.

(3) f is *bijective* or is a *bijection* if it is both injective and surjective. If such a bijection f exists from A to B, we say A and B are in *bijective correspondence*.

(4) f has a *left inverse* if there is a function $g : B \to A$ such that $g \circ f : A \to A$ is the identity map on A, i.e. $(g \circ f)(a) = a$, for all $a \in A$.

(5) f has a *right inverse* if there is a function $h : B \to A$ such that $f \circ h : B \to B$ is the identity map on B.

Proposition 1. Let $f : A \to B$.

(1) The map f is injective if and only if f has a left inverse.

(2) The map f is surjective if and only if f has a right inverse.

(3) The map f is a bijection if and only if there exists $g : B \to A$ such that $f \circ g$ is the identity map on B and $g \circ f$ is the identity map on A.

(4) If A and B are finite sets with the same number of elements (i.e. $|A| = |B|$), then $f : A \to B$ is bijective if and only if f is injective if and only if f is surjective.

Proof: Exercise.

In the situation of part (3) of the proposition above the map g is necessarily unique and we shall say g is the *2-sided inverse* (or simply the *inverse*) of f.

A *permutation* of a set A is simply a bijection from A to itself.

If $A \subseteq B$ and $f : B \to C$, we denote the *restriction* of f to A by $f|_A$. When the domain we are considering is understood we shall occasionally denote $f|_A$ again simply as f even though these are formally different functions (their domains are different).

If $A \subseteq B$ and $g : A \to C$ and there is a function $f : B \to C$ such that $f|_A = g$, we shall say f is an *extension* of g to B (such a map f need not exist nor be unique).

Let A be a nonempty set.

(1) A *binary relation* on a set A is a subset R of $A \times A$ and we write $a \sim b$ if $(a, b) \in R$.

(2) The relation $\sim$ on A is said to be:

(a) *reflexive* if $a \sim a$, for all $a \in A$,

(b) *symmetric* if $a \sim b$ implies $b \sim a$ for all $a, b \in A$,

(c) *transitive* if $a \sim b$ and $b \sim c$ implies $a \sim c$ for all $a, b, c \in A$.

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

(3) If $\sim$ defines an equivalence relation on A, then the *equivalence class* of $a \in A$ is defined to be $\{x \in A \mid x \sim a\}$. Elements of the equivalence class of a are said to be *equivalent* to a. If C is an equivalence class, any element of C is called a *representative* of the class C.

(4) A *partition* of A is any collection $\{A_i \mid i \in I\}$ of nonempty subsets of A (I some indexing set) such that

(a) $A = \cup_{i \in I} A_i$, and

(b) $A_i \cap A_j = \emptyset$, for all $i, j \in I$ with $i \neq j$,

i.e. A is the disjoint union of the sets in the partition.

The notions of an equivalence relation on A and a partition of A are the same:

Proposition 2. Let A be a nonempty set.

(1) If $\sim$ defines an equivalence relation on A then the set of equivalence classes of $\sim$ form a partition of A.

(2) If $\{A_i \mid i \in I\}$ is a partition of A then there is an equivalence relation on A whose equivalence classes are precisely the sets A_i, $i \in I$.

Proof: Omitted.

Finally, we shall assume the reader is familiar with proofs by induction.

EXERCISES

In Exercises 1 to 4 let $\mathcal{A}$ be the set of 2×2 matrices with real number entries. Recall that matrix multiplication is defined by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} p & q \\ r & s \end{pmatrix} = \begin{pmatrix} ap + br & aq + bs \\ cp + dr & cq + ds \end{pmatrix}.$$

Let

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

and let

$$\mathcal{B} = \{X \in \mathcal{A} \mid MX = XM\}.$$

1. Determine which of the following elements of $\mathcal{A}$ lie in $\mathcal{B}$:

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

2. Prove that if $P, Q \in \mathcal{B}$, then $P + Q \in \mathcal{B}$ (where $+$ denotes the usual sum of two matrices).
3. Prove that if $P, Q \in \mathcal{B}$, then $P \cdot Q \in \mathcal{B}$ (where $\cdot$ denotes the usual product of two matrices).
4. Find conditions on p, q, r, s which determine precisely when $\begin{pmatrix} p & q \\ r & s \end{pmatrix} \in \mathcal{B}$.
5. Determine whether the following functions f are well defined:
   (a) $f : \mathbb{Q} \to \mathbb{Z}$ defined by $f(a/b) = a$.
   (b) $f : \mathbb{Q} \to \mathbb{Q}$ defined by $f(a/b) = a^2/b^2$.
6. Determine whether the function $f : \mathbb{R}^+ \to \mathbb{Z}$ defined by mapping a real number r to the first digit to the right of the decimal point in a decimal expansion of r is well defined.
7. Let $f : A \to B$ be a surjective map of sets. Prove that the relation

$$a \sim b \text{ if and only if } f(a) = f(b)$$

is an equivalence relation whose equivalence classes are the fibers of f.

0.2 PROPERTIES OF THE INTEGERS

The following properties of the integers $\mathbb{Z}$ (many familiar from elementary arithmetic) will be proved in a more general context in the ring theory of Chapter 8, but it will be necessary to use them in Part I (of course, none of the ring theory proofs of these properties will rely on the group theory).

(1) (Well Ordering of $\mathbb{Z}$) If A is any nonempty subset of $\mathbb{Z}^+$, there is some element $m \in A$ such that $m \leq a$, for all $a \in A$ (m is called a *minimal element* of A).

(2) If $a, b \in \mathbb{Z}$ with $a \neq 0$, we say a *divides* b if there is an element $c \in \mathbb{Z}$ such that $b = ac$. In this case we write $a \mid b$; if a does not divide b we write $a \nmid b$.

(3) If $a, b \in \mathbb{Z} - \{0\}$, there is a unique positive integer d, called the *greatest common divisor of a and b* (or g.c.d. of a and b), satisfying:
   (a) $d \mid a$ and $d \mid b$ (so d is a common divisor of a and b), and
   (b) if $e \mid a$ and $e \mid b$, then $e \mid d$ (so d is the greatest such divisor).
   The g.c.d. of a and b will be denoted by (a, b). If $(a, b) = 1$, we say that a and b are *relatively prime*.

(4) If $a, b \in \mathbb{Z} - \{0\}$, there is a unique positive integer l, called the *least common multiple of a and b* (or l.c.m. of a and b), satisfying:

(a) $a \mid l$ and $b \mid l$ (so l is a common multiple of a and b), and
(b) if $a \mid m$ and $b \mid m$, then $l \mid m$ (so l is the least such multiple).
The connection between the greatest common divisor d and the least common multiple l of two integers a and b is given by $dl = ab$.

(5) The *Division Algorithm* : if $a, b \in \mathbb{Z} - \{0\}$, then there exist unique $q, r \in \mathbb{Z}$ such that

$$a = qb + r \quad \text{and} \quad 0 \leq r < |b|,$$

where q is the *quotient* and r the *remainder*. This is the usual "long division" familiar from elementary arithmetic.

(6) The *Euclidean Algorithm* is an important procedure which produces a greatest common divisor of two integers a and b by iterating the Division Algorithm: if $a, b \in \mathbb{Z} - \{0\}$, then we obtain a sequence of quotients and remainders

$$\begin{aligned} a &= q_0 b + r_0 && (0) \\ b &= q_1 r_0 + r_1 && (1) \\ r_0 &= q_2 r_1 + r_2 && (2) \\ r_1 &= q_3 r_2 + r_3 && (3) \\ &\vdots \\ r_{n-2} &= q_n r_{n-1} + r_n && (n) \\ r_{n-1} &= q_{n+1} r_n && (n+1) \end{aligned}$$

where r_n is the last nonzero remainder. Such an r_n exists since $|b| > |r_0| > |r_1| > \cdots > |r_n|$ is a decreasing sequence of strictly positive integers if the remainders are nonzero and such a sequence cannot continue indefinitely. Then r_n is the g.c.d. (a, b) of a and b.

Example

Suppose $a = 57970$ and $b = 10353$. Then applying the Euclidean Algorithm we obtain:

$$\begin{aligned} 57970 &= (5)10353 + 6205 \\ 10353 &= (1)6205 + 4148 \\ 6205 &= (1)4148 + 2057 \\ 4148 &= (2)2057 + 34 \\ 2057 &= (60)34 + 17 \\ 34 &= (2)17 \end{aligned}$$

which shows that $(57970, 10353) = 17$.

(7) One consequence of the Euclidean Algorithm which we shall use regularly is the following: if $a, b \in \mathbb{Z} - \{0\}$, then there exist $x, y \in \mathbb{Z}$ such that

$$(a, b) = ax + by$$

that is, *the g.c.d. of a and b is a $\mathbb{Z}$-linear combination of a and b.* This follows by recursively writing the element r_n in the Euclidean Algorithm in terms of the

previous remainders (namely, use equation (n) above to solve for $r_n = r_{n-2} - q_n r_{n-1}$ in terms of the remainders r_{n-1} and r_{n-2}, then use equation $(n-1)$ to write r_n in terms of the remainders r_{n-2} and r_{n-3}, etc., eventually writing r_n in terms of a and b).

Example

Suppose $a = 57970$ and $b = 10353$, whose greatest common divisor we computed above to be 17. From the fifth equation (the next to last equation) in the Euclidean Algorithm applied to these two integers we solve for their greatest common divisor: $17 = 2057 - (60)34$. The fourth equation then shows that $34 = 4148 - (2)2057$, so substituting this expression for the previous remainder 34 gives the equation $17 = 2057 - (60)[4148 - (2)2057]$, i.e. $17 = (121)2057 - (60)4148$. Solving the third equation for 2057 and substituting gives $17 = (121)[6205 - (1)4148] - (60)4148 = (121)6205 - (181)4148$. Using the second equation to solve for 4148 and then the first equation to solve for 6205 we finally obtain

$$17 = (302)57970 - (1691)10353$$

as can easily be checked directly. Hence the equation $ax + by = (a, b)$ for the greatest common divisor of a and b in this example has the solution $x = 302$ and $y = -1691$. Note that it is relatively unlikely that this relation would have been found simply by guessing.

The integers x and y in (7) above are not unique. In the example with $a = 57970$ and $b = 10353$ we determined one solution to be $x = 302$ and $y = -1691$, for instance, and it is relatively simple to check that $x = -307$ and $y = 1719$ also satisfy $57970x + 10353y = 17$. The general solution for x and y is known (cf. the exercises below and in Chapter 8).

(8) An element p of $\mathbb{Z}^+$ is called a *prime* if $p > 1$ and the only positive divisors of p are 1 and p (initially, the word prime will refer only to positive integers). An integer $n > 1$ which is not prime is called *composite*. For example, 2,3,5,7,11,13,17,19, ... are primes and 4,6,8,9,10,12,14,15,16,18, ... are composite.

An important property of primes (which in fact can be used to *define* the primes (cf. Exercise 3)) is the following: if p is a prime and $p \mid ab$, for some $a, b \in \mathbb{Z}$, then either $p \mid a$ or $p \mid b$.

(9) The *Fundamental Theorem of Arithmetic* says: if $n \in \mathbb{Z}$, $n > 1$, then n can be factored uniquely into the product of primes, i.e. there are distinct primes $p_1, p_2, \ldots, p_s$ and positive integers $\alpha_1, \alpha_2, \ldots, \alpha_s$ such that

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \ldots p_s^{\alpha_s}.$$

This factorization is unique in the sense that if $q_1, q_2, \ldots, q_t$ are any distinct primes and $\beta_1, \beta_2, \ldots, \beta_t$ positive integers such that

$$n = q_1^{\beta_1} q_2^{\beta_2} \ldots q_t^{\beta_t},$$

then $s = t$ and if we arrange the two sets of primes in increasing order, then $q_i = p_i$ and $\alpha_i = \beta_i$, $1 \le i \le s$. For example, $n = 1852423848 = 2^3 3^2 11^2 19^3 31$ and this decomposition into the product of primes is unique.

Suppose the positive integers a and b are expressed as products of prime powers:

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_s^{\alpha_s}, \quad b = p_1^{\beta_1} p_2^{\beta_2} \cdots p_s^{\beta_s}$$

where $p_1, p_2, \ldots, p_s$ are distinct and the exponents are ≥ 0 (we allow the exponents to be 0 here so that the products are taken over the same set of primes – the exponent will be 0 if that prime is not actually a divisor). Then the greatest common divisor of a and b is

$$(a, b) = p_1^{\min(\alpha_1, \beta_1)} p_2^{\min(\alpha_2, \beta_2)} \cdots p_s^{\min(\alpha_s, \beta_s)}$$

(and the least common multiple is obtained by instead taking the maximum of the α_i and β_i instead of the minimum).

Example

In the example above, $a = 57970$ and $b = 10353$ can be factored as $a = 2 \cdot 5 \cdot 11 \cdot 17 \cdot 31$ and $b = 3 \cdot 7 \cdot 17 \cdot 29$, from which we can immediately conclude that their greatest common divisor is 17. Note, however, that for large integers it is extremely difficult to determine their prime factorizations (several common codes in current use are based on this difficulty, in fact), so that this is not an effective method to determine greatest common divisors in general. The Euclidean Algorithm will produce greatest common divisors quite rapidly without the need for the prime factorization of a and b.

(10) The *Euler φ–function* is defined as follows: for $n \in \mathbb{Z}^+$ let $\varphi(n)$ be the number of positive integers $a \leq n$ with a relatively prime to n, i.e. $(a, n) = 1$. For example, $\varphi(12) = 4$ since 1, 5, 7 and 11 are the only positive integers less than or equal to 12 which have no factors in common with 12. Similarly, $\varphi(1) = 1$, $\varphi(2) = 1$, $\varphi(3) = 2$, $\varphi(4) = 2$, $\varphi(5) = 4$, $\varphi(6) = 2$, etc. For primes p, $\varphi(p) = p - 1$, and, more generally, for all $a \geq 1$ we have the formula

$$\varphi(p^a) = p^a - p^{a-1} = p^{a-1}(p - 1).$$

The function φ is *multiplicative* in the sense that

$$\varphi(ab) = \varphi(a)\varphi(b) \qquad \text{if } (a, b) = 1$$

(note that it is important here that a and b be relatively prime). Together with the formula above this gives a general formula for the values of φ : if $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_s^{\alpha_s}$, then

$$\begin{aligned}\varphi(n) &= \varphi(p_1^{\alpha_1})\varphi(p_2^{\alpha_2}) \cdots \varphi(p_s^{\alpha_s}) \\ &= p_1^{\alpha_1 - 1}(p_1 - 1)p_2^{\alpha_2 - 1}(p_2 - 1) \cdots p_s^{\alpha_s - 1}(p_s - 1).\end{aligned}$$

For example, $\varphi(12) = \varphi(2^2)\varphi(3) = 2^1(2 - 1)3^0(3 - 1) = 4$. The reader should note that we shall use the letter φ for many different functions throughout the text so when we want this letter to denote Euler's function we shall be careful to indicate this explicitly.

EXERCISES

1. For each of the following pairs of integers a and b, determine their greatest common divisor, their least common multiple, and write their greatest common divisor in the form $ax + by$ for some integers x and y.

(a) $a = 20, b = 13$.
(b) $a = 69, b = 372$.
(c) $a = 792, b = 275$.
(d) $a = 11391, b = 5673$.
(e) $a = 1761, b = 1567$.
(f) $a = 507885, b = 60808$.

2. Prove that if the integer k divides the integers a and b then k divides $as + bt$ for every pair of integers s and t.
3. Prove that if n is composite then there are integers a and b such that n divides ab but n does not divide either a or b.
4. Let a, b and N be fixed integers with a and b nonzero and let $d = (a, b)$ be the greatest common divisor of a and b. Suppose x_0 and y_0 are particular solutions to $ax + by = N$ (i.e. $ax_0 + by_0 = N$). Prove for any integer t that the integers

$$x = x_0 + \frac{b}{d}t \quad \text{and} \quad y = y_0 - \frac{a}{d}t$$

are also solutions to $ax + by = N$ (this is in fact the general solution).
5. Determine the value $\varphi(n)$ for each integer $n \leq 30$ where φ denotes the Euler φ-function.
6. Prove the Well Ordering Property of $\mathbb{Z}$ by induction and prove the minimal element is unique.
7. If p is a prime prove that there do not exist nonzero integers a and b such that $a^2 = pb^2$ (i.e. $\sqrt{p}$ is not a rational number).
8. Let p be a prime, $n \in \mathbb{Z}^+$. Find a formula for the largest power of p which divides $n! = n(n-1)(n-2)\ldots 2 \cdot 1$ (it involves the greatest integer function).
9. Write a computer program to determine the greatest common divisor (a, b) of two integers a and b and to express (a, b) in the form $ax + by$ for some integers x and y.
10. Prove for any given positive integer N there exist only finitely many integers n with $\varphi(n) = N$ where φ denotes Euler's φ-function. Conclude in particular that $\varphi(n)$ tends to infinity as n tends to infinity.
11. Prove that if d divides n then $\varphi(d)$ divides $\varphi(n)$ where φ denotes Euler's φ-function.

0.3 $\mathbb{Z}/n\mathbb{Z}$: THE INTEGERS MODULO *n*

Let n be a fixed positive integer. Define a relation on $\mathbb{Z}$ by

$$a \sim b \text{ if and only if } n \mid (b - a).$$

Clearly $a \sim a$, and $a \sim b$ implies $b \sim a$ for any integers a and b, so this relation is trivially reflexive and symmetric. If $a \sim b$ and $b \sim c$ then n divides $a - b$ and n divides $b - c$ so n also divides the sum of these two integers, i.e. n divides $(a-b)+(b-c) = a-c$, so $a \sim c$ and the relation is transitive. Hence this is an equivalence relation. Write $a \equiv b \pmod{n}$ (read: a is *congruent* to b mod n) if $a \sim b$. For any $k \in \mathbb{Z}$ we shall denote the equivalence class of a by $\bar{a}$ - this is called the *congruence class* or *residue class* of a mod n and consists of the integers which differ from a by an integral multiple of n, i.e.

$$\begin{aligned} \bar{a} &= \{a + kn \mid k \in \mathbb{Z}\} \\ &= \{a, a \pm n, a \pm 2n, a \pm 3n, \ldots\}. \end{aligned}$$

There are precisely n distinct equivalence classes mod n, namely

$$\bar{0}, \bar{1}, \bar{2}, \ldots, \overline{n-1}$$

determined by the possible remainders after division by n and these residue classes partition the integers $\mathbb{Z}$. The set of equivalence classes under this equivalence relation will be denoted by $\mathbb{Z}/n\mathbb{Z}$ and called the *integers modulo* n (or the *integers mod* n). The motivation for this notation will become clearer when we discuss quotient groups and quotient rings. Note that for different n's the equivalence relation and equivalence classes are different so we shall always be careful to fix n first before using the bar notation. The process of finding the equivalence class mod n of some integer a is often referred to as *reducing* a *mod* n. This terminology also frequently refers to finding the smallest nonnegative integer congruent to a mod n (the *least residue* of a mod n).

We can define an addition and a multiplication for the elements of $\mathbb{Z}/n\mathbb{Z}$, defining *modular arithmetic* as follows: for $\bar{a}, \bar{b} \in \mathbb{Z}/n\mathbb{Z}$, define their sum and product by

$$\bar{a} + \bar{b} = \overline{a+b} \qquad \text{and} \qquad \bar{a} \cdot \bar{b} = \overline{ab}.$$

What this means is the following: given any two elements $\bar{a}$ and $\bar{b}$ in $\mathbb{Z}/n\mathbb{Z}$, to compute their sum (respectively, their product) take *any representative* integer a in the *class* $\bar{a}$ and *any representative* integer b in the *class* $\bar{b}$ and add (respectively, multiply) the integers a and b as usual in $\mathbb{Z}$ and then take the equivalence class containing the result. The following Theorem 3 asserts that this is well defined, i.e. does not depend on the choice of representatives taken for the elements $\bar{a}$ and $\bar{b}$ of $\mathbb{Z}/n\mathbb{Z}$.

Example

Suppose $n = 12$ and consider $\mathbb{Z}/12\mathbb{Z}$, which consists of the twelve residue classes

$$\bar{0}, \bar{1}, \bar{2}, \ldots, \overline{11}$$

determined by the twelve possible remainders of an integer after division by 12. The elements in the residue class $\bar{5}$, for example, are the integers which leave a remainder of 5 when divided by 12 (the integers *congruent to* 5 mod 12). Any integer congruent to 5 mod 12 (such as 5, 17, 29, ... or $-7, -19, \ldots$) will serve as a representative for the residue class $\bar{5}$. Note that $\mathbb{Z}/12\mathbb{Z}$ consists of the twelve *elements* above (and each of these elements of $\mathbb{Z}/12\mathbb{Z}$ consists of an infinite number of usual integers).

Suppose now that $\bar{a} = \bar{5}$ and $\bar{b} = \bar{8}$. The most obvious representative for $\bar{a}$ is the integer 5 and similarly 8 is the most obvious representative for $\bar{b}$. Using *these* representatives for the residue classes we obtain $\bar{5} + \bar{8} = \overline{13} = \bar{1}$ since 13 and 1 lie in the same class modulo $n = 12$. Had we instead taken the representative 17, say, for $\bar{a}$ (note that 5 and 17 do lie in the same residue class modulo 12) and the representative -28, say, for $\bar{b}$, we would obtain $\bar{5} + \bar{8} = \overline{(17-28)} = \overline{-11} = \bar{1}$ and as we mentioned the result does not depend on the choice of representatives chosen. The product of these two classes is $\bar{a} \cdot \bar{b} = \overline{5 \cdot 8} = \overline{40} = \bar{4}$, also independent of the representatives chosen.

Theorem 3. The operations of addition and multiplication on $\mathbb{Z}/n\mathbb{Z}$ defined above are both well defined, that is, they do not depend on the choices of representatives for the classes involved. More precisely, if $a_1, a_2 \in \mathbb{Z}$ and $b_1, b_2 \in \mathbb{Z}$ with

$\overline{a_1} = \overline{b_1}$ and $\overline{a_2} = \overline{b_2}$, then $\overline{a_1 + a_2} = \overline{b_1 + b_2}$ and $\overline{a_1 a_2} = \overline{b_1 b_2}$, i.e. if

$$a_1 \equiv b_1 \pmod{n} \quad \text{and} \quad a_2 \equiv b_2 \pmod{n}$$

then

$$a_1 + a_2 \equiv b_1 + b_2 \pmod{n} \quad \text{and} \quad a_1 a_2 \equiv b_1 b_2 \pmod{n}.$$

Proof: Suppose $a_1 \equiv b_1 \pmod{n}$, i.e. $a_1 - b_1$ is divisible by n. Then $a_1 = b_1 + sn$ for some integer s. Similarly, $a_2 \equiv b_2 \pmod{n}$ means $a_2 = b_2 + tn$ for some integer t. Then $a_1 + a_2 = (b_1 + b_2) + (s + t)n$ so that $a_1 + a_2 \equiv b_1 + b_2 \pmod{n}$, which shows that the sum of the residue classes is independent of the representatives chosen. Similarly, $a_1 a_2 = (b_1 + sn)(b_2 + tn) = b_1 b_2 + (b_1 t + b_2 s + st)n$ shows that $a_1 a_2 \equiv b_1 b_2 \pmod{n}$ and so the product of the residue classes is also independent of the representatives chosen, completing the proof.

We shall see later that the process of adding equivalence classes by adding their representatives is a special case of a more general construction (the construction of a *quotient*). This notion of adding equivalence classes is already a familiar one in the context of adding rational numbers : each rational number a/b is really a class of expressions: $a/b = 2a/2b = -3a/-3b$ etc. and we often change representatives (for instance, take common denominators) in order to add two fractions (for example $1/2 + 1/3$ is computed by taking instead the equivalent representatives $3/6$ for $1/2$ and $2/6$ for $1/3$ to obtain $1/2 + 1/3 = 3/6 + 2/6 = 5/6$). The notion of modular arithmetic is also familiar: to find the hour of day after adding or subtracting some number of hours we reduce mod 12 and find the least residue.

It is important to be able to think of the equivalence classes of some equivalence relation as *elements* which can be manipulated (as we do, for example, with fractions) rather than as sets. Consistent with this attitude, we shall frequently denote the elements of $\mathbb{Z}/n\mathbb{Z}$ simply by $\{0, 1, \ldots, n-1\}$ where addition and multiplication are *reduced mod* n. It is important to remember, however, that the elements of $\mathbb{Z}/n\mathbb{Z}$ are *not* integers, but rather collections of usual integers, and the arithmetic is quite different. For example, $5 + 8$ is not 1 in the integers $\mathbb{Z}$ as it was in the example of $\mathbb{Z}/12\mathbb{Z}$ above.

The fact that one can define arithmetic in $\mathbb{Z}/n\mathbb{Z}$ has many important applications in elementary number theory. As one simple example we compute the last two digits in the number 2^{1000}. First observe that the last two digits give the remainder of 2^{1000} after we divide by 100 so we are interested in the residue class mod 100 containing 2^{1000}. We compute $2^{10} = 1024 \equiv 24 \pmod{100}$, so then $2^{20} = (2^{10})^2 \equiv 24^2 = 576 \equiv 76 \pmod{100}$. Then $2^{40} = (2^{20})^2 \equiv 76^2 = 5776 \equiv 76 \pmod{100}$. Similarly $2^{80} \equiv 2^{160} \equiv 2^{320} \equiv 2^{640} \equiv 76 \pmod{100}$. Finally, $2^{1000} = 2^{640}2^{320}2^{40} \equiv 76 \cdot 76 \cdot 76 \equiv 76 \pmod{100}$ so the final two digits are 76.

An important subset of $\mathbb{Z}/n\mathbb{Z}$ consists of the collection of residue classes which have a multiplicative inverse in $\mathbb{Z}/n\mathbb{Z}$:

$$(\mathbb{Z}/n\mathbb{Z})^\times = \{\bar{a} \in \mathbb{Z}/n\mathbb{Z} \mid \text{there exists } \bar{c} \in \mathbb{Z}/n\mathbb{Z} \text{ with } \bar{a} \cdot \bar{c} = \bar{1}\}.$$

Some of the following exercises outline a proof that $(\mathbb{Z}/n\mathbb{Z})^\times$ is also the collection of residue classes whose representatives are relatively prime to n:

Proposition 4. $(\mathbb{Z}/n\mathbb{Z})^\times = \{\bar{a} \in \mathbb{Z}/n\mathbb{Z} \mid (a, n) = 1\}$.

It is easy to see that if *any* representative of $\bar{a}$ is relatively prime to n then *all* representatives are relatively prime to n so that the set on the right in the proposition is well defined.

Example

For $n = 9$ we obtain $(\mathbb{Z}/9\mathbb{Z})^\times = \{1, 2, 4, 5, 7, 8\}$ from the proposition. The multiplicative inverses of these elements are $\{1, 5, 7, 2, 4, 8\}$, respectively.

If a is an integer relatively prime to n then the Euclidean Algorithm produces integers x and y satisfying $ax + ny = 1$, hence $ax \equiv 1 \pmod{n}$, so that $\bar{x}$ is the multiplicative inverse of $\bar{a}$ in $\mathbb{Z}/n\mathbb{Z}$. This gives an efficient method for computing multiplicative inverses in $\mathbb{Z}/n\mathbb{Z}$.

Example

Suppose $n = 60$ and $a = 17$. Applying the Euclidean Algorithm we obtain

$$\begin{aligned} 60 &= (3)17 + 9 \\ 17 &= (1)9 + 8 \\ 9 &= (1)8 + 1 \end{aligned}$$

so that a and n are relatively prime, and $(-7)17 + (2)60 = 1$. Hence $-\overline{7} = \overline{53}$ is the multiplicative inverse of $\overline{17}$ in $\mathbb{Z}/60\mathbb{Z}$.

EXERCISES

1. Write down explicitly all the elements in the residue classes of $\mathbb{Z}/18\mathbb{Z}$.
2. Prove that the distinct equivalence classes in $\mathbb{Z}/n\mathbb{Z}$ are precisely $\bar{0}, \bar{1}, \bar{2}, \ldots, \overline{n-1}$ (use the Division Algorithm).
3. Prove that if $a = a_n 10^n + a_{n-1} 10^{n-1} + \cdots + a_1 10 + a_0$ is any positive integer then $a \equiv a_n + a_{n-1} + \cdots + a_1 + a_0 \pmod{9}$ (note that this is the usual arithmetic rule that the remainder after division by 9 is the same as the sum of the decimal digits mod 9 – in particular an integer is divisible by 9 if and only if the sum of its digits is divisible by 9) [note that $10 \equiv 1 \pmod{9}$].
4. Compute the remainder when 37^{100} is divided by 29.
5. Compute the last two digits of 9^{1500}.
6. Prove that the squares of the elements in $\mathbb{Z}/4\mathbb{Z}$ are just $\bar{0}$ and $\bar{1}$.
7. Prove for any integers a and b that $a^2 + b^2$ never leaves a remainder of 3 when divided by 4 (use the previous exercise).
8. Prove that the equation $a^2 + b^2 = 3c^2$ has no solutions in nonzero integers a, b and c. [Consider the equation mod 4 as in the previous two exercises and show that a, b and c would all have to be divisible by 2. Then each of a^2, b^2 and c^2 has a factor of 4 and by dividing through by 4 show that there would be a smaller set of solutions to the original equation. Iterate to reach a contradiction.]
9. Prove that the square of any odd integer always leaves a remainder of 1 when divided by 8.

10. Prove that the number of elements of $(\mathbb{Z}/n\mathbb{Z})^\times$ is $\varphi(n)$ where φ denotes the Euler φ-function.

11. Prove that if $\bar{a}, \bar{b} \in (\mathbb{Z}/n\mathbb{Z})^\times$, then $\bar{a} \cdot \bar{b} \in (\mathbb{Z}/n\mathbb{Z})^\times$.

12. Let $n \in \mathbb{Z}$, $n > 1$, and let $a \in \mathbb{Z}$ with $1 \leq a \leq n$. Prove if a and n are not relatively prime, there exists an integer b with $1 \leq b < n$ such that $ab \equiv 0 \pmod{n}$ and deduce that there cannot be an integer c such that $ac \equiv 1 \pmod{n}$.

13. Let $n \in \mathbb{Z}$, $n > 1$, and let $a \in \mathbb{Z}$ with $1 \leq a \leq n$. Prove that if a and n are relatively prime then there is an integer c such that $ac \equiv 1 \pmod{n}$ [use the fact that the g.c.d. of two integers is a $\mathbb{Z}$-linear combination of the integers].

14. Conclude from the previous two exercises that $(\mathbb{Z}/n\mathbb{Z})^\times$ is the set of elements $\bar{a}$ of $\mathbb{Z}/n\mathbb{Z}$ with $(a, n) = 1$ and hence prove Proposition 4. Verify this directly in the case $n = 12$.

15. For each of the following pairs of integers a and n, show that a is relatively prime to n and determine the multiplicative inverse of $\bar{a}$ in $\mathbb{Z}/n\mathbb{Z}$.
    (a) $a = 13, n = 20$.
    (b) $a = 69, n = 89$.
    (c) $a = 1891, n = 3797$.
    (d) $a = 6003722857, n = 77695236973$. [The Euclidean Algorithm requires only 3 steps for these integers.]

16. Write a computer program to add and multiply mod n, for any n given as input. The output of these operations should be the least residues of the sums and products of two integers. Also include the feature that if $(a, n) = 1$, an integer c between 1 and $n - 1$ such that $\bar{a} \cdot \bar{c} = \bar{1}$ may be printed on request. (Your program should not, of course, simply quote "mod" functions already built into many systems).

Part I

GROUP THEORY

The modern treatment of abstract algebra begins with the disarmingly simple abstract definition of a *group*. This simple definition quickly leads to difficult questions involving the structure of such objects. There are many specific examples of groups and the power of the abstract point of view becomes apparent when results for *all* of these examples are obtained by proving a *single* result for the abstract group.

The notion of a group did not simply spring into existence, however, but is rather the culmination of a long period of mathematical investigation, the first formal definition of an abstract group in the form in which we use it appearing in 1882.[1] The definition of an abstract group has its origins in extremely old problems in algebraic equations, number theory, and geometry, and arose because very similar techniques were found to be applicable in a variety of situations. As Otto Hölder (1859–1937) observed, one of the essential characteristics of mathematics is that after applying a certain algorithm or method of proof one then considers the scope and limits of the method. As a result, properties possessed by a number of interesting objects are frequently abstracted and the question raised: can one determine *all* the objects possessing these properties? Attempting to answer such a question also frequently adds considerable understanding of the original objects under consideration. It is in this fashion that the definition of an abstract group evolved into what is, for us, the starting point of abstract algebra.

We illustrate with a few of the disparate situations in which the ideas later formalized into the notion of an abstract group were used:

1. In number theory the very object of study, the set of integers, is an example of a group. Consider for example what we refer to as "Euler's Theorem" (cf. Exercise 22 of Section 3.2), one extremely simple example of which is that a^{40} has last two digits 01 if a is any integer not divisible by 2 nor by 5. This was proved in 1761 by Leonhard Euler (1707–1783) using "group-theoretic" ideas of Joseph Louis Lagrange (1736–1813), long before the first formal definition of a group. From our perspective, one now proves "Lagrange's Theorem" (cf. Theorem 8 of Section 3.2), applying these

[1] For most of the historical comments below, see the excellent book *A History of Algebra*, by B. L. van der Waerden, Springer-Verlag, 1980 and the references there, particularly *The Genesis of the Abstract Group Concept: A Contribution to the History of the Origin of Abstract Group Theory* (translated from the German by Abe Shenitzer), by H. Wussing, MIT Press, 1984. See also *Number Theory, An Approach Through History from Hammurapai to Legendre*, by A. Weil, Birkhäuser, 1984.

techniques abstracted to an arbitary group, and then *recovers* Euler's Theorem (and many others) as a *special case*.

2. Investigations into the question of rational solutions to algebraic equations of the form $y^2 = x^3 - 2x$ (there are infinitely many, for example $(0,0)$, $(-1,1)$, $(2,2)$, $(9/4,-21/8)$, $(-1/169, 239/2197)$) showed that connecting any two solutions by a straight line and computing the intersection of this line with the curve $y^2 = x^3 - 2x$ produces another solution. Such "Diophantine equations", among others, were considered by Pierre de Fermat (1601–1655) (this one was solved by him in 1644), by Euler, by Lagrange around 1777, and others. In 1730 Euler raised the question of determining the indefinite integral $\int dx/\sqrt{1-x^4}$ of the "lemniscatic differential" $dx/\sqrt{1-x^4}$, used in determining the arc length along an ellipse (the question had also been considered by Gottfried Wilhelm Leibniz (1646–1716) and Johannes Bernoulli (1667–1748)). In 1752 Euler proved a "multiplication formula" for such elliptic integrals (using ideas of G.C. di Fagnano (1682–1766), received by Euler in 1751), which shows how two elliptic integrals give rise to a third, bringing into existence the theory of elliptic functions in analysis. In 1834 Carl Gustav Jacob Jacobi (1804–1851) observed that the work of Euler on solving certain Diophantine equations amounted to writing the multiplication formula for certain elliptic integrals. Today the curve above is referred to as an "elliptic curve" and these questions are viewed as two different aspects of the same thing — the fact that this geometric operation on points can be used to give the set of points on an elliptic curve the structure of a group. The study of the "arithmetic" of these groups is an active area of current research.[2]

3. By 1824 it was known that there are formulas giving the roots of quadratic, cubic and quartic equations (extending the familiar quadratic formula for the roots of $ax^2 + bx + c = 0$). In 1824, however, Niels Henrik Abel (1802–1829) proved that such a formula for the roots of a quintic is impossible (cf. Corollary 40 of Section 14.7). The proof is based on the idea of examining what happens when the roots are permuted amongst themselves (for example, interchanging two of the roots). The collection of such permutations has the structure of a group (called, naturally enough, a "permutation group"). This idea culminated in the beautiful work of Evariste Galois (1811–1832) in 1830–32, working with explicit groups of "substitutions". Today this work is referred to as Galois Theory (and is the subject of the fourth part of this text). Similar explicit groups were being used in geometry as collections of geometric transformations (translations, reflections, etc.) by Arthur Cayley (1821–1895) around 1850, Camille Jordan (1838–1922) around 1867, Felix Klein (1849–1925) around 1870, etc., and the application of groups to geometry is still extremely active in current research into the structure of 3-space, 4-space, etc. The same group arising in the study of the solvability of the quintic arises in the study of the rigid motions of an icosahedron in geometry and in the study of elliptic functions in analysis.

The precursors of today's abstract group can be traced back many years, even before the groups of "substitutions" of Galois. The formal definition of an abstract group which is our starting point appeared in 1882 in the work of Walter Dyck (1856–

[2] See *The Arithmetic of Elliptic Curves* by J. Silverman, Springer-Verlag, 1986.

1934), an assistant to Felix Klein, and also in the work of Heinrich Weber (1842–1913) in the same year.

It is frequently the case in mathematics research to find specific application of an idea before having that idea extracted and presented as an item of interest in its own right (for example, Galois used the notion of a "quotient group" implicitly in his investigations in 1830 and the definition of an abstract quotient group is due to Hölder in 1889). It is important to realize, with or without the historical context, that the reason the abstract definitions are made is because it is useful to isolate specific characteristics and consider what structure is imposed on an object having these characteristics. The notion of the structure of an algebraic object (which is made more precise by the concept of an isomorphism — which considers when two apparently different objects are in some sense the same) is a major theme which will recur throughout the text.

1

Introduction to Groups

1.1 BASIC AXIOMS AND EXAMPLES

In this section the basic algebraic structure to be studied in Part I is introduced and some examples are given.

Definition

(1) A *binary operation* $\star$ on a set G is a function $\star : G \times G \to G$. For any $a, b \in G$ we shall write $a \star b$ for $\star(a, b)$.

(2) A binary operation $\star$ on a set G is *associative* if for all $a, b, c \in G$, $a \star (b \star c) = (a \star b) \star c$.

(3) If $\star$ is a binary operation on a set G we say elements a and b of G *commute* if $a \star b = b \star a$. We say $\star$ (or G) is *commutative* if for all $a, b \in G$, $a \star b = b \star a$.

Examples

(1) $+$ (usual addition) is a commutative binary operation on $\mathbb{Z}$ (or on $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ respectively).

(2) $\times$ (usual multiplication) is a commutative binary operation on $\mathbb{Z}$ (or on $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ respectively).

(3) $-$ (usual subtraction) is a noncommutative binary operation on $\mathbb{Z}$, where $-(a, b) = a - b$. The map $a \mapsto -a$ is not a binary operation (not binary).

(4) $-$ is not a binary operation on $\mathbb{Z}^+$ (nor $\mathbb{Q}^+$, $\mathbb{R}^+$) because for $a, b \in \mathbb{Z}^+$ with $a < b$, $a - b \notin \mathbb{Z}^+$, that is, $-$ does not map $\mathbb{Z}^+ \times \mathbb{Z}^+$ into $\mathbb{Z}^+$.

(5) Taking the vector cross–product of two vectors in 3-space $\mathbb{R}^3$ is a binary operation which is not associative and not commutative.

Suppose that $\star$ is a binary operation on a set G and H is a subset of G. If the restriction of $\star$ to H is a binary operation on H, i.e. for all $a, b \in H$, $a \star b \in H$, then H is said to be *closed* under $\star$. Observe that if $\star$ is an associative (respectively, commutative) binary operation on G and $\star$ restricted to some subset H of G is a binary operation on H, then $\star$ is automatically associative (respectively, commutative) on H as well.

Definition

(1) A *group* is an ordered pair $(G, \star)$ where G is a set and $\star$ is a binary operation on G satisfying the following axioms:

(i) $(a \star b) \star c = a \star (b \star c)$, for all $a, b, c \in G$ (i.e. $\star$ is associative),

(ii) there exists an element e in G such that $a \star e = e \star a = a$, for all $a \in G$ (e is called an *identity* of G),

(iii) for each $a \in G$ there is an element a^{-1} of G such that $a \star a^{-1} = a^{-1} \star a = e$ (a^{-1} is called an *inverse* of a).

(2) The group $(G, \star)$ is called *abelian* (or *commutative*) if $a \star b = b \star a$ for all $a, b \in G$.

We shall immediately become less formal and say G is a group under $\star$ if $(G, \star)$ is a group (or just G is a group when the operation $\star$ is clear from the context). Also, we say G is a *finite group* if in addition G is a finite set. Note that axiom (ii) ensures that a group is always nonempty.

Examples

(1) $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are groups under $+$ with $e = 0$ and $a^{-1} = -a$, for all a.

(2) $\mathbb{Q} - \{0\}$, $\mathbb{R} - \{0\}$, $\mathbb{C} - \{0\}$, $\mathbb{Q}^+$, $\mathbb{R}^+$ are groups under $\times$ with $e = 1$ and $a^{-1} = \frac{1}{a}$, for all a. Note however that $\mathbb{Z} - \{0\}$ is *not* a group under $\times$ because although $\times$ is an associative binary operation on $\mathbb{Z} - \{0\}$, the element 2 (for instance) does not have an inverse in $\mathbb{Z} - \{0\}$.

We have glossed over the fact that the associative law holds in these familiar examples. For $\mathbb{Z}$ under $+$ this is a consequence of the axiom of associativity for addition of natural numbers. The associative law for $\mathbb{Q}$ under $+$ follows from the associative law for $\mathbb{Z}$ — a proof of this will be outlined later when we rigorously construct $\mathbb{Q}$ from $\mathbb{Z}$ (cf. Section 7.5). The associative laws for $\mathbb{R}$ and, in turn, $\mathbb{C}$ under $+$ are proved in elementary analysis courses when $\mathbb{R}$ is constructed by completing $\mathbb{Q}$ — ultimately, associativity is again a consequence of associativity for $\mathbb{Z}$. The associative axiom for multiplication may be established via a similar development, starting first with $\mathbb{Z}$. Since $\mathbb{R}$ and $\mathbb{C}$ will be used largely for illustrative purposes and we shall not construct $\mathbb{R}$ from $\mathbb{Q}$ (although we shall construct $\mathbb{C}$ from $\mathbb{R}$) we shall take the associative laws (under $+$ and $\times$) for $\mathbb{R}$ and $\mathbb{C}$ as given.

Examples (continued)

(3) The axioms for a vector space V include those axioms which specify that $(V, +)$ is an abelian group (the operation $+$ is called vector addition). Thus any vector space such as $\mathbb{R}^n$ is, in particular, an additive group.

(4) For $n \in \mathbb{Z}^+$, $\mathbb{Z}/n\mathbb{Z}$ is an abelian group under the operation $+$ of addition of residue classes as described in Chapter 0. We shall prove in Chapter 3 (in a more general context) that this binary operation $+$ is well defined and associative; for now we take this for granted. The identity in this group is the element $\bar{0}$ and for each $\bar{a} \in \mathbb{Z}/n\mathbb{Z}$, the inverse of $\bar{a}$ is $\overline{-a}$. Henceforth, when we talk about the group $\mathbb{Z}/n\mathbb{Z}$ it will be understood that the group operation is addition of classes mod n.

(5) For $n \in \mathbb{Z}^+$, the set $(\mathbb{Z}/n\mathbb{Z})^\times$ of equivalence classes $\bar{a}$ which have multiplicative inverses mod n is an abelian group under *multiplication* of residue classes as described in Chapter 0. Again, we shall take for granted (for the moment) that this operation is well defined and associative. The identity of this group is the element $\bar{1}$ and, by definition of $(\mathbb{Z}/n\mathbb{Z})^\times$, each element has a multiplicative inverse. Henceforth, when

we talk about the group $(\mathbb{Z}/n\mathbb{Z})^\times$ it will be understood that the group operation is multiplication of classes mod n.

(6) If $(A, \star)$ and $(B, \diamond)$ are groups, we can form a new group $A \times B$, called their *direct product*, whose elements are those in the Cartesian product $A \times B = \{(a, b) \mid a \in A,\ b \in B\}$ and whose operation is defined componentwise:

$$(a_1, b_1)(a_2, b_2) = (a_1 \star a_2, b_1 \diamond b_2).$$

For example, if we take $A = B = \mathbb{R}$ (both operations addition), $\mathbb{R} \times \mathbb{R}$ is the familiar Euclidean plane. The proof that the direct product of two groups is again a group is left as a straightforward exercise (later) — the proof that each group axiom holds in $A \times B$ is a consequence of that axiom holding in both A and B together with the fact that the operation in $A \times B$ is defined componentwise.

There should be no confusion between the groups $\mathbb{Z}/n\mathbb{Z}$ (under addition) and $(\mathbb{Z}/n\mathbb{Z})^\times$ (under multiplication), even though the latter is a subset of the former — the superscript $\times$ will always indicate that the operation is multiplication.

Before continuing with more elaborate examples we prove two basic results which in particular enable us to talk about *the* identity and *the* inverse of an element.

Proposition 1 If G is a group under the operation $\star$, then

(1) the identity of G is unique

(2) for each $a \in G$, a^{-1} is uniquely determined

(3) $(a^{-1})^{-1} = a$ for all $a \in G$

(4) $(a \star b)^{-1} = (b^{-1}) \star (a^{-1})$

(5) for any $a_1, a_2, \ldots, a_n \in G$ the value of $a_1 \star a_2 \star \cdots \star a_n$ is independent of how the expression is bracketed (this is called the *generalized associative law*).

Proof: (1) If f and g are both identities, then by axiom (ii) of the definition of a group $f \star g = f$ (take $a = f$ and $e = g$). By the same axiom $f \star g = g$ (take $a = g$ and $e = f$). Thus $f = g$, and the identity is unique.

(2) Assume b and c are both inverses of a and let e be the identity of G. By axiom (iii), $a \star b = e$ and $c \star a = e$. Thus

$$\begin{aligned} c &= c \star e && \text{(definition of } e \text{ — axiom (ii))} \\ &= c \star (a \star b) && \text{(since } e = a \star b\text{)} \\ &= (c \star a) \star b && \text{(associative law)} \\ &= e \star b && \text{(since } e = c \star a\text{)} \\ &= b && \text{(axiom (ii)).} \end{aligned}$$

(3) To show $(a^{-1})^{-1} = a$ is exactly the problem of showing a is the inverse of a^{-1} (since by part (2) a has a unique inverse). Reading the definition of a^{-1}, with the roles of a and a^{-1} mentally interchanged shows that a satisfies the defining property for the inverse of a^{-1}, hence a is the inverse of a^{-1}.

(4) Let $c = (a \star b)^{-1}$ so by definition of c, $(a \star b) \star c = e$. By the associative law

$$a \star (b \star c) = e.$$

Multiply both sides on the left by a^{-1} to get

$$a^{-1} \star (a \star (b \star c)) = a^{-1} \star e.$$

The associative law on the left hand side and the definition of e on the right give

$$(a^{-1} \star a) \star (b \star c) = a^{-1}$$

so

$$e \star (b \star c) = a^{-1}$$

hence

$$b \star c = a^{-1}.$$

Now multiply both sides on the left by b^{-1} and simplify similarly:

$$\begin{aligned} b^{-1} \star (b \star c) &= b^{-1} \star a^{-1} \\ (b^{-1} \star b) \star c &= b^{-1} \star a^{-1} \\ e \star c &= b^{-1} \star a^{-1} \\ c &= b^{-1} \star a^{-1}, \end{aligned}$$

as claimed.

(5) This is left as a good exercise using induction on n. First show the result is true for $n = 1, 2$, and 3. Next assume for any $k < n$ that any bracketing of a product of k elements, $b_1 \star b_2 \star \cdots \star b_k$ can be reduced (without altering the value of the product) to an expression of the form

$$b_1 \star (b_2 \star (b_3 \star (\cdots \star b_k)) \ldots).$$

Now argue that any bracketing of the product $a_1 \star a_2 \star \cdots \star a_n$ must break into 2 subproducts, say $(a_1 \star a_2 \star \cdots \star a_k) \star (a_{k+1} \star a_{k+2} \star \cdots \star a_n)$, where each subproduct is bracketed in some fashion. Apply the induction assumption to each of these two subproducts and finally reduce the result to the form $a_1 \star (a_2 \star (a_3 \star (\cdots \star a_n)) \ldots)$ to complete the induction.

Note that throughout the proof of Proposition 1 we were careful not to change the *order* of any products (unless permitted by axioms (ii) and (iii)) since G may be non-abelian.

Notation:

(1) For an abstract group G it is tiresome to keep writing the operation $\star$ throughout our calculations. Henceforth (except when necessary) our abstract groups G, H, etc. will always be written with the operation as $\cdot$ and $a \cdot b$ will always be written as ab. In view of the generalized associative law, products of three or more group elements will not be bracketed (although the operation is still a binary operation). Finally, for an abstract group G (operation $\cdot$) we shall denote the identity of G by 1.

(2) For any group G (operation $\cdot$ implied) and $x \in G$ and $n \in \mathbb{Z}^+$ since the product $xx \cdots x$ (n terms) does not depend on how it is bracketed, we shall denote it by x^n. Denote $x^{-1}x^{-1} \cdots x^{-1}$ (n terms) by x^{-n}. Let $x^0 = 1$, the identity of G.

This new notation is pleasantly concise. Of course, when we are dealing with specific groups, we shall use the natural (given) operation. For example, when the operation is $+$, the identity will be denoted by 0 and for any element a, the inverse a^{-1} will be written $-a$ and $a + a + \cdots + a$ ($n > 0$ terms) will be written na; $-a - a \cdots - a$ (n terms) will be written $-na$ and $0a = 0$.

Proposition 2. Let G be a group and let $a, b \in G$. The equations $ax = b$ and $ya = b$ have unique solutions for $x, y \in G$. In particular, the left and right cancellation laws hold in G, i.e.

(1) if $au = av$, then $u = v$, and

(2) if $ub = vb$, then $u = v$.

Proof: We can solve $ax = b$ by multiplying both sides on the left by a^{-1} and simplifying to get $x = a^{-1}b$. The uniqueness of x follows because a^{-1} is unique. Similarly, if $ya = b$, $y = ba^{-1}$. If $au = av$, multiply both sides on the left by a^{-1} and simplify to get $u = v$. Similarly, the right cancellation law holds.

One consequence of Proposition 2 is that if a is any element of G and for some $b \in G$, $ab = e$ or $ba = e$, then $b = a^{-1}$, i.e. we do not have to show both equations hold. Also, if for some $b \in G$, $ab = a$ (or $ba = a$), then b must be the identity of G, i.e. we do not have to check $bx = xb = x$ for all $x \in G$.

Definition. For G a group and $x \in G$ define the *order* of x to be the smallest positive integer n such that $x^n = 1$, and denote this integer by $|x|$. In this case x is said to be of order n. If no positive power of x is the identity, the order of x is defined to be infinity and x is said to be of infinite order.

The symbol for the order of x should not be confused with the absolute value symbol (when $G \subseteq \mathbb{R}$ we shall be careful to distinguish the two). It may seem injudicious to choose the same symbol for order of an element as the one used to denote the cardinality (or order) of a set, however, we shall see that the order of an element in a group is the same as the cardinality of the set of all its (distinct) powers so the two uses of the the word "order" are naturally related.

Examples

(1) An element of a group has order 1 if and only if it is the identity.

(2) In the additive groups $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ or $\mathbb{C}$ every nonzero (i.e. nonidentity) element has infinite order.

(3) In the multiplicative groups $\mathbb{R} - \{0\}$ or $\mathbb{Q} - \{0\}$ the element -1 has order 2 and all other nonidentity elements have infinite order.

(4) In the additive group $\mathbb{Z}/9\mathbb{Z}$ the element $\bar{6}$ has order 3, since $\bar{6} \neq \bar{0}$, $\bar{6} + \bar{6} = \overline{12} = \bar{3} \neq \bar{0}$, but $\bar{6} + \bar{6} + \bar{6} = \overline{18} = \bar{0}$, the identity in this group. Recall that in an *additive*

group the powers of an element are the integer multiples of the element. Similarly, the order of the element $\bar{5}$ is 9, since this is the smallest positive multiple of 5 that is divisible by 9.

(5) In the multiplicative group $(\mathbb{Z}/7\mathbb{Z})^\times$, the powers of the element $\bar{2}$ are $\bar{2}, \bar{4}, \bar{8} = 1$, the identity in this group, so $\bar{2}$ has order 3. Similarly, the element $\bar{3}$ has order 6, since 6 is the smallest positive power of 3 that is divisible by 7.

Definition. Let $G = \{g_1, g_2, \ldots, g_n\}$ be a finite group with $g_1 = 1$. The *multiplication table* or *group table* of G is the $n \times n$ matrix whose i, j entry is the group element $g_i g_j$.

For a finite group the multiplication table contains, in some sense, all the information about the group. Computationally, however, it is an unwieldly object (being of size the square of the group order) and visually it is not a very useful object for determining properties of the group. One might think of a group table as the analog of having a table of all the distances between pairs of cities in the country. Such a table is useful and, in essence, captures all the distance relationships, yet a map (better yet, a map with all the distances labelled on it) is a much easier tool to work with. Part of our initial development of the theory of groups (finite groups in particular) is directed towards a more conceptual way of visualizing the internal structure of groups.

EXERCISES

Let G be a group.

1. Determine which of the following binary operations are associative:
   (a) the operation $\star$ on $\mathbb{Z}$ defined by $a \star b = a - b$
   (b) the operation $\star$ on $\mathbb{R}$ defined by $a \star b = a + b + ab$
   (c) the operation $\star$ on $\mathbb{Q}$ defined by $a \star b = \dfrac{a+b}{5}$
   (d) the operation $\star$ on $\mathbb{Z} \times \mathbb{Z}$ defined by $(a, b) \star (c, d) = (ad + bc, bd)$
   (e) the operation $\star$ on $\mathbb{Q} - \{0\}$ defined by $a \star b = \dfrac{a}{b}$.
2. Decide which of the binary operations in the preceding exercise are commutative.
3. Prove that addition of residue classes in $\mathbb{Z}/n\mathbb{Z}$ is associative (you may assume it is well defined).
4. Prove that multiplication of residue classes in $\mathbb{Z}/n\mathbb{Z}$ is associative (you may assume it is well defined).
5. Prove for all $n > 1$ that $\mathbb{Z}/n\mathbb{Z}$ is not a group under multiplication of residue classes.
6. Determine which of the following sets are groups under addition:
   (a) the set of rational numbers (including $0 = 0/1$) in lowest terms whose denominators are odd
   (b) the set of rational numbers (including $0 = 0/1$) in lowest terms whose denominators are even
   (c) the set of rational numbers of absolute value < 1
   (d) the set of rational numbers of absolute value ≥ 1 together with 0
   (e) the set of rational numbers with denominators at most 2
   (f) the set of rational numbers with denominators at most 3.

7. Let $G = \{x \in \mathbb{R} \mid 0 \leq x < 1\}$ and for $x, y \in G$ let $x \star y$ be the fractional part of $x + y$ (i.e. $x \star y = x + y - [x + y]$ where $[a]$ is the greatest integer less than or equal to a). Prove that $\star$ is a well defined binary operation on G and that G is an abelian group under $\star$ (called the *real numbers mod 1*).

8. Let $G = \{z \in \mathbb{C} \mid z^n = 1 \text{ for some } n \in \mathbb{Z}^+\}$.
   (a) Prove that G is a group under multiplication (called the group of *roots of unity* in $\mathbb{C}$).
   (b) Prove that G is not a group under addition.

9. Let $G = \{a + b\sqrt{2} \in \mathbb{R} \mid a, b \in \mathbb{Q}\}$.
   (a) Prove that G is a group under addition.
   (b) Prove that the nonzero elements of G are a group under multiplication. ["Rationalize the denominators" to find multiplicative inverses.]

10. Prove that a finite group is abelian if and only if its group table is a symmetric matrix.

11. Find the orders of each element of the additive group $\mathbb{Z}/12\mathbb{Z}$.

12. Find the orders of the following elements of the multiplicative group $(\mathbb{Z}/12\mathbb{Z})^\times$: $\overline{1}$, $\overline{-1}$, $\overline{5}$, $\overline{7}$, $\overline{-7}$, $\overline{13}$.

13. Find the orders of the following elements of the additive group $\mathbb{Z}/36\mathbb{Z}$: $\overline{1}$, $\overline{2}$, $\overline{6}$, $\overline{9}$, $\overline{10}$, $\overline{12}$, $\overline{-1}$, $\overline{-10}$, $\overline{-18}$.

14. Find the orders of the following elements of the multiplicative group $(\mathbb{Z}/36\mathbb{Z})^\times$: $\overline{1}$, $\overline{-1}$, $\overline{5}$, $\overline{13}$, $\overline{-13}$, $\overline{17}$.

15. Prove that $(a_1 a_2 \ldots a_n)^{-1} = a_n^{-1} a_{n-1}^{-1} \ldots a_1^{-1}$ for all $a_1, a_2, \ldots, a_n \in G$.

16. Let x be an element of G. Prove that $x^2 = 1$ if and only if $|x|$ is either 1 or 2.

17. Let x be an element of G. Prove that if $|x| = n$ for some positive integer n then $x^{-1} = x^{n-1}$.

18. Let x and y be elements of G. Prove that $xy = yx$ if and only if $y^{-1}xy = x$ if and only if $x^{-1}y^{-1}xy = 1$.

19. Let $x \in G$ and let $a, b \in \mathbb{Z}^+$.
   (a) Prove that $x^{a+b} = x^a x^b$ and $(x^a)^b = x^{ab}$.
   (b) Prove that $(x^a)^{-1} = x^{-a}$.
   (c) Establish part (a) for arbitrary integers a and b (positive, negative or zero).

20. For x an element in G show that x and x^{-1} have the same order.

21. Let G be a finite group and let x be an element of G of order n. Prove that if n is odd, then $x = (x^2)^k$ for some k.

22. If x and g are elements of the group G, prove that $|x| = |g^{-1}xg|$. Deduce that $|ab| = |ba|$ for all $a, b \in G$.

23. Suppose $x \in G$ and $|x| = n < \infty$. If $n = st$ for some positive integers s and t, prove that $|x^s| = t$.

24. If a and b are *commuting* elements of G, prove that $(ab)^n = a^n b^n$ for all $n \in \mathbb{Z}$. [Do this by induction for positive n first.]

25. Prove that if $x^2 = 1$ for all $x \in G$ then G is abelian.

26. Assume H is a nonempty subset of $(G, \star)$ which is closed under the binary operation on G and is closed under inverses, i.e. for all h and $k \in H$, hk and $h^{-1} \in H$. Prove that H is a group under the operation $\star$ restricted to H (such a subset H is called a *subgroup* of G).

27. Prove that if x is an element of the group G then $\{x^n \mid n \in \mathbb{Z}\}$ is a subgroup (cf. the preceding exercise) of G (called the *cyclic subgroup* of G generated by x).

28. Let $(A, \star)$ and $(B, \diamond)$ be groups and let $A \times B$ be their direct product (as defined in Example 6). Verify all the group axioms for $A \times B$:

(a) prove that the associative law holds: for all $(a_i, b_i) \in A \times B$, $i = 1, 2, 3$ $(a_1, b_1)[(a_2, b_2)(a_3, b_3)] = [(a_1, b_1)(a_2, b_2)](a_3, b_3)$.

(b) prove that $(1, 1)$ is the identity of $A \times B$, and

(c) prove that the inverse of (a, b) is (a^{-1}, b^{-1}).

29. Prove that $A \times B$ is an abelian group if and only if both A and B are abelian.

30. Prove that the elements $(a, 1)$ and $(1, b)$ of $A \times B$ commute and deduce that the order of (a, b) is the least common multiple of $|a|$ and $|b|$.

31. Prove that any finite group G of even order contains an element of order 2. [Let $t(G) = \{g \in G \mid g \neq g^{-1}\}$. Show that $t(G)$ has an even number of elements and every nonidentity element of $G - t(G)$ has order 2.]

32. If x is an element of finite order n in G, prove that the elements $1, x, x^2, \ldots, x^{n-1}$ are all distinct. Deduce that $|x| \leq |G|$.

33. Let x be an element of finite order n in G.

(a) Prove that if n is odd then $x^i \neq x^{-i}$ for all $i = 1, 2, \ldots, n-1$.

(b) Prove that if $n = 2k$ is even and $x^i = x^{-i}$ then $i \equiv k \bmod n$.

34. If x is an element of infinite order in G, prove that the elements x^n, $n \in \mathbb{Z}$ are all distinct.

35. If x is an element of finite order n in G, use the Division Algorithm to show that *any* integral power of x equals one of the elements in the set $\{1, x, x^2, \ldots, x^{n-1}\}$ (so these are all the distinct elements of the cyclic subgroup (cf. Exercise 27 above) of G generated by x).

36. Assume $G = \{1, a, b, c\}$ is a group of order 4 with identity 1. Assume also that G has no elements of order 4 (so by Exercise 32, every element has order ≤ 3). Use the cancellation laws to show that there is a unique group table for G. Deduce that G is abelian.

1.2 DIHEDRAL GROUPS

An important family of examples of groups is the class of groups whose elements are symmetries of geometric objects. The simplest subclass is when the geometric objects are regular planar figures.

For each $n \in \mathbb{Z}^+$, $n \geq 3$ let D_{2n} be the set of symmetries of a regular n-gon, where a symmetry is any rigid motion of the n-gon which can be effected by taking a copy of the n-gon, moving this copy in any fashion in 3-space and then placing the copy back on the original n-gon so it exactly covers it. More precisely, we can describe the symmetries by first choosing a labelling of the n vertices, for example as shown in the following figure.

Then each symmetry s can be described uniquely by the corresponding permutation σ of $\{1, 2, 3, \ldots, n\}$ where if the symmetry s puts vertex i in the place where vertex j was originally, then σ is the permutation sending i to j. For instance, if s is a rotation of $2\pi/n$ radians clockwise about the center of the n-gon, then σ is the permutation sending i to $i + 1$, $1 \leq i \leq n - 1$, and $\sigma(n) = 1$. Now make D_{2n} into a group by defining st for $s, t \in D_{2n}$ to be the symmetry obtained by first applying t then s to the n-gon (note that we are viewing symmetries as functions on the n-gon, so st is just function composition — read as usual from right to left). If s, t effect the permutations σ, τ, respectively on the vertices, then st effects $\sigma \circ \tau$. The binary oper-

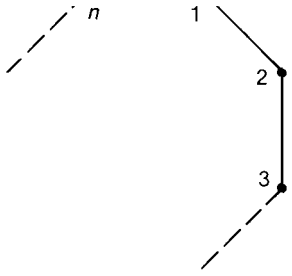


Fig. 1

ation on D_{2n} is associative since composition of functions is associative. The identity of D_{2n} is the identity symmetry (which leaves all vertices fixed), denoted by 1, and the inverse of $s \in D_{2n}$ is the symmetry which reverses all rigid motions of s (so if s effects permutation σ on the vertices, s^{-1} effects σ^{-1}). In the next paragraph we show

$$|D_{2n}| = 2n$$

and so D_{2n} is called the *dihedral group of order* $2n$. In some texts this group is written D_n; however, D_{2n} (where the subscript gives the order of the group rather than the number of vertices) is more common in the group theory literature.

To find the order $|D_{2n}|$ observe that given any vertex i, there is a symmetry which sends vertex 1 into position i. Since vertex 2 is adjacent to vertex 1, vertex 2 must end up in position $i+1$ or $i-1$ (where $n+1$ is 1 and $1-1$ is n, i.e. the integers labelling the vertices are read mod n). Moreover, by following the first symmetry by a reflection about the line through vertex i and the center of the n-gon one sees that vertex 2 can be sent to either position $i+1$ or $i-1$ by some symmetry. Thus there are $n \cdot 2$ positions the ordered pair of vertices 1, 2 may be sent to upon applying symmetries. Since symmetries are rigid motions one sees that once the position of the ordered pair of vertices 1, 2 has been specified, the action of the symmetry on all remaining vertices is completely determined. Thus there are exactly $2n$ symmetries of a regular n-gon. We can, moreover, explicitly exhibit $2n$ symmetries. These symmetries are: the n rotations about the center through $2\pi i/n$ radian, $0 \leq i \leq n-1$, and the n reflections through the n lines of symmetry (if n is odd, each symmetry line passes through a vertex and

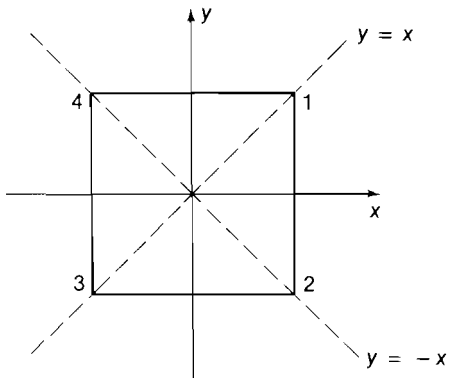


Fig. 2

the mid-point of the opposite side; if n is even, there are $n/2$ lines of symmetry which pass through 2 opposite vertices and $n/2$ which perpendicularly bisect two opposite sides). For example, if $n = 4$ and we draw a square at the origin in an x, y plane, the lines of symmetry are the lines $x = 0$ (y-axis), $y = 0$ (x-axis), $y = x$ and $y = -x$ (note that "reflection" through the origin is not a reflection but a rotation of π radians).

Since dihedral groups will be used extensively as an example throughout the text we fix some notation and mention some calculations which will simplify future computations and assist in viewing D_{2n} as an abstract group (rather than having to return to the geometric setting at every instance). Fix a regular n-gon centered at the origin in an x, y plane and label the vertices consecutively from 1 to n in a clockwise manner. Let r be the rotation clockwise about the origin through $2\pi/n$ radian. Let s be the reflection about the line of symmetry through vertex 1 and the origin (we use the same letters for each n, but the context will always make n clear). We leave the details of the following calculations as an exercise (for the most part we shall be working with D_6 and D_8, so the reader may wish to try these exercises for $n = 3$ and $n = 4$ first):

(1) $1, r, r^2, \ldots, r^{n-1}$ are all distinct and $r^n = 1$, so $|r| = n$.

(2) $|s| = 2$.

(3) $s \neq r^i$ for any i.

(4) $sr^i \neq sr^j$, for all $0 \leq i, j \leq n-1$ with $i \neq j$, so

$$D_{2n} = \{1, r, r^2, \ldots, r^{n-1}, s, sr, sr^2, \ldots, sr^{n-1}\}$$

i.e. each element can be written *uniquely* in the form $s^k r^i$ for some $k = 0$ or 1 and $0 \leq i \leq n-1$.

(5) $sr = r^{-1}s$. [First work out what permutation s effects on $\{1, 2, \ldots, n\}$ and then work out separately what each side in this equation does to vertices 1 and 2.] This shows in particular that r and s do not commute so that D_{2n} is non-abelian.

(6) $sr^i = r^{-i}s$, for all $0 \leq i \leq n$. [Proceed by induction on i and use the fact that $sr^{i+1} = (sr^i)r$ together with the preceding calculation.] This indicates how to commute s with powers of r.

Having done these calculations, we now observe that the complete multiplication table of D_{2n} can be written in terms r and s alone, that is, all the elements of D_{2n} have a (unique) representation in the form $s^k r^i$, $k = 0$ or 1 and $0 \leq i \leq n-1$, and any product of two elements in this form can be reduced to another in the same form using only (1), (2) and (6) (reducing all exponents mod n). For example, if $n = 12$,

$$(sr^9)(sr^6) = s(r^9 s)r^6 = s(sr^{-9})r^6 = s^2 r^{-9+6} = r^{-3} = r^9.$$

The following general remarks will be discussed more thoroughly when we introduce free groups (Section 6.3); they provide a simple way of computing in D_{2n}. In any group G a subset, S, of elements with the property that every element of G can be written as a product of elements of S is called a set of *generators* of G and we shall indicate this by $G = \langle S \rangle$ and say G is *generated by* S. Any equations in G that these generators satisfy are called *relations* in G. In D_{2n} we found a set of generators,

$\{r, s\}$, and a set of relations ((1), (2) and (5) above, for instance) which completely determine when any two expressions in terms of r and s are equal. Equivalently, this says any relation between these two generators can be deduced from (1), (2) and (5) alone. If some group G is generated by a subset S and there is some set of relations, say $R_1, R_2, .., R_m$ (here each R_i is some equation in the elements from $S \cup \{1\}$) such that any relation among the elements of S can be deduced from these, we shall call these generators and relations a *presentation* of G and write

$$G = \langle S \mid R_1, R_2, \ldots, R_m \rangle.$$

We shall henceforth do all our calculations in D_{2n} in terms of the presentation computed above and display the requisite information as

$$D_{2n} = \langle r, s \mid r^n = s^2 = 1, \quad rs = sr^{-1} \rangle.$$

EXERCISES

Let $D_{2n} = \langle r, s \mid r^n = s^2 = 1,\ rs = sr^{-1} \rangle$.

1. Compute the order of each of the elements in the following groups:
   (a) D_6 (b) D_8 (c) D_{10}.
2. Use the generators and relations above to show that if x is any element of D_{2n} which is not a power of r, then $rx = xr^{-1}$.
3. Use the generators and relations above to show that every element of D_{2n} which is not a power of r has order 2. Deduce that D_{2n} is generated by the two elements s and sr, both of which have order 2.
4. If $n = 2k$ is even and $n \geq 4$, show that $z = r^k$ is an element of order 2 which commutes with all elements of D_{2n}. Show also that z is the only nonidentity element of D_{2n} which commutes with all elements of D_{2n}. [cf. Exercise 33 of Section 1.]
5. If n is odd and $n \geq 3$, show that the identity is the only element of D_{2n} which commutes with all elements of D_{2n}. [cf. Exercise 33 of Section 1.]
6. Let x and y be elements of order 2 in any group G. Prove that if $t = xy$ then $tx = xt^{-1}$ (so that if $n = |xy| < \infty$ then x, t satisfy the same relations in G as s, r do in D_{2n}).
7. Show that $\langle a, b \mid a^2 = b^2 = (ab)^n = 1 \rangle$ gives a presentation for D_{2n} in terms of the two generators $a = s$ and $b = sr$ of order 2 computed in Exercise 3 above. [Show that the relations for r and s follow from the relations for a and b and, conversely, the relations for a and b follow from those for r and s.]
8. Find the order of the cyclic subgroup of D_{2n} generated by r (cf. Exercise 27 of Section 1).

In each of Exercises 9 to 13 you can find the order of the group of rigid motions in $\mathbb{R}^3$ (also called the group of rotations) of the given Platonic solid by following the proof for the order of D_{2n}: find the number of positions to which an adjacent pair of vertices can be sent. Alternatively, you can find the number of places to which a given face may be sent and, once a face is fixed, the number of positions to which a vertex on that face may be sent.

9. Let G be the group of rigid motions in $\mathbb{R}^3$ of a tetrahedron. Show that $|G| = 12$.
10. Let G be the group of rigid motions in $\mathbb{R}^3$ of a cube. Show that $|G| = 24$.
11. Let G be the group of rigid motions in $\mathbb{R}^3$ of an octahedron. Show that $|G| = 24$.
12. Let G be the group of rigid motions in $\mathbb{R}^3$ of a dodecahedron. Show that $|G| = 60$.
13. Let G be the group of rigid motions in $\mathbb{R}^3$ of an icosahedron. Show that $|G| = 60$.

1.3 SYMMETRIC GROUPS

Let Ω be any nonempty set and let S_Ω be the set of all bijections from Ω to itself (i.e. the set of all permutations of Ω). The set S_Ω is a group under function composition: $\circ$. Note that $\circ$ is a binary operation on S_Ω since if $\sigma : \Omega \to \Omega$ and $\tau : \Omega \to \Omega$ are both bijections, then $\sigma \circ \tau$ is also a bijection from Ω to Ω . Since function composition is associative in general, $\circ$ is associative. The identity of S_Ω is the permutation 1 defined by $1(a) = a$, for all $a \in \Omega$. For every permutation σ there is a (2-sided) inverse function, $\sigma^{-1} : \Omega \to \Omega$ satisfying $\sigma \circ \sigma^{-1} = \sigma^{-1} \circ \sigma = 1$. Thus, all the group axioms hold for $(S_\Omega, \circ)$. This group is called the *symmetric group on the set* Ω . It is important to recognize that the elements of S_Ω are the *permutations* of Ω , not the elements of Ω itself.

In the special case when $\Omega = \{1, 2, 3, \ldots, n\}$, the symmetric group on Ω is denoted S_n, the *symmetric group of degree* n.[1] The group S_n will play an important role throughout the text both as a group of considerable interest in its own right and as a means of illustrating and motivating the general theory.

First we show that the order of S_n is $n!$. The permutations of $\{1, 2, 3, \ldots, n\}$ are precisely the injective functions of this set to itself because it is finite (Proposition 0.1) and we can count the number of injective functions. An injective function σ can send the number 1 to any of the n elements of $\{1, 2, 3, \ldots, n\}$; $\sigma(2)$ can then be any one of the elements of this set except $\sigma(1)$ (so there are $n - 1$ choices for $\sigma(2)$); $\sigma(3)$ can be any element except $\sigma(1)$ or $\sigma(2)$ (so there are $n - 2$ choices for $\sigma(3)$), and so on. Thus there are precisely $n \cdot (n - 1) \cdot (n - 2) \ldots 2 \cdot 1 = n!$ possible injective functions from $\{1, 2, 3, \ldots, n\}$ to itself. Hence there are precisely $n!$ permutations of $\{1, 2, 3, \ldots, n\}$ so there are precisely $n!$ elements in S_n.

We now describe an efficient notation for writing elements σ of S_n which we shall use throughout the text and which is called the *cycle decomposition*.

A *cycle* is a string of integers which represents the element of S_n which cyclically permutes these integers (and fixes all other integers). The cycle $(a_1\, a_2\, \ldots\, a_m)$ is the permutation which sends a_i to a_{i+1}, $1 \le i \le m - 1$ and sends a_m to a_1. For example $(2\,1\,3)$ is the permutation which maps 2 to 1, 1 to 3 and 3 to 2. In general, for each $\sigma \in S_n$ the numbers from 1 to n will be rearranged and grouped into k cycles of the form

$$(a_1\, a_2 \ldots a_{m_1})(a_{m_1+1}\, a_{m_1+2} \ldots a_{m_2}) \ldots (a_{m_{k-1}+1}\, a_{m_{k-1}+2} \ldots a_{m_k})$$

from which the action of σ on any number from 1 to n can easily be read, as follows. For any $x \in \{1, 2, 3, \ldots, n\}$ first locate x in the above expression. If x is not followed immediately by a right parenthesis (i.e. x is not at the right end of one of the k cycles), then $\sigma(x)$ is the integer appearing immediately to the right of x. If x is followed by a right parenthesis, then $\sigma(x)$ is the number which is at the start of the cycle ending with x (i.e. if $x = a_{m_i}$, for some i, then $\sigma(x) = a_{m_{i-1}+1}$ (where m_0 is taken to be 0)). We can represent this description of σ by:

$$\begin{tikzcd} a_1 \arrow[r] & a_2 \arrow[r] & \cdots \arrow[r] & a_{m_1} \arrow[lll, bend left] \end{tikzcd}$$

[1] We shall see in Section 6 that the structure of S_Ω depends only on the cardinality of Ω , not on the particular elements of Ω itself, so if Ω is any finite set with n elements, then S_Ω "looks like" S_n.

$$\circlearrowright a_{m_1+1} \to a_{m_1+2} \to \cdots \to a_{m_2} \circlearrowleft$$

$$\vdots$$

$$\circlearrowright a_{m_{k-1}+1} \to a_{m_{k-1}+2} \to \cdots \to a_{m_k} \circlearrowleft$$

The product of all the cycles is called the *cycle decomposition* of σ .

We now give an algorithm for computing the cycle decomposition of an element σ of S_n and work through the algorithm with a specific permutation. We defer the proof of this algorithm and full analysis of the uniqueness aspects of the cycle decomposition until Chapter 4.

Let $n = 13$ and let $\sigma \in S_{13}$ be defined by

$$\begin{array}{lllll} \sigma(1) = 12, & \sigma(2) = 13, & \sigma(3) = 3, & \sigma(4) = 1, & \sigma(5) = 11, \\ \sigma(6) = 9, & \sigma(7) = 5, & \sigma(8) = 10, & \sigma(9) = 6, & \sigma(10) = 4, \\ \sigma(11) = 7, & \sigma(12) = 8, & \sigma(13) = 2. & & \end{array}$$

Cycle Decomposition Algorithm

Method	Example
To start a new cycle pick the smallest element of $\{1,2,\ldots,n\}$ which has not yet appeared in a previous cycle—call it a (if you are just starting, $a = 1$); begin the new cycle: $(a$	$(1$
Read off $\sigma(a)$ from the given description of σ—call it b. If $b = a$, close the cycle with a right parenthesis (without writing b down); this completes a cycle—return to step 1. If $b \neq a$, write b next to a in this cycle: $(a\ b$	$\sigma(1) = 12 = b$, $12 \neq 1$ so write: $(1\ 12$
Read off $\sigma(b)$ from the given description of σ—call it c. If $c = a$, close the cycle with a right parenthesis to complete the cycle—return to step 1. If $c \neq a$, write c next to b in this cycle: $(a\ b\ c$ Repeat this step using the number c as the new value for b until the cycle closes.	$\sigma(12) = 8$, $8 \neq 1$ so continue the cycle as: $(1\ 12\ 8$

Naturally this process stops when all the numbers from $\{1, 2, \ldots, n\}$ have appeared in some cycle. For the particular σ in the example this gives:

$$\sigma = (1\ 12\ 8\ 10\ 4)(2\ 13)(3)(5\ 11\ 7)(6\ 9)$$

The *length* of a cycle is the number of integers which appear in it. A cycle of length t is called a *t-cycle*. Two cycles are called *disjoint* if they have no numbers in common. Thus the element σ above is the product of 5 (pairwise) disjoint cycles: a 5-cycle, a 2-cycle, a 1-cycle, a 3-cycle, and another 2-cycle.

Henceforth we adopt the convention that 1-cycles will not be written. Thus if some integer, i, does not appear in the cycle decomposition of a permutation τ it is understood that $\tau(i) = i$, i.e. that τ fixes i. The identity permutation of S_n has cycle

decomposition $(1)(2)\ldots(n)$ and will be written simply as 1. Hence the final step of the algorithm is:

Final Step: Remove all cycles of length 1	

The cycle decomposition for the particular σ in the example is therefore

$$\sigma = (1\ 12\ 8\ 10\ 4)(2\ 13)(5\ 11\ 7)(6\ 9)$$

This convention has the advantage that the cycle decomposition of an element τ of S_n is also the cycle decomposition of the permutation in S_m for $m \geq n$ which acts as τ on $\{1, 2, 3, \ldots, n\}$ and fixes each element of $\{n+1, n+2, \ldots, m\}$. Thus, for example, $(1\ 2)$ is the permutation which interchanges 1 and 2 and fixes all larger integers whether viewed in S_2, S_3 or S_4, etc.

As another example, the 6 elements of S_3 have the following cycle decompositions:

The Group S_3

Values of σ_i	Cycle decomposition of σ_i
$\sigma_1(1) = 1, \sigma_1(2) = 2, \sigma_1(3) = 3$	1
$\sigma_2(1) = 1, \sigma_2(2) = 3, \sigma_2(3) = 2$	$(2\ 3)$
$\sigma_3(1) = 3, \sigma_3(2) = 2, \sigma_3(3) = 1$	$(1\ 3)$
$\sigma_4(1) = 2, \sigma_4(2) = 1, \sigma_4(3) = 3$	$(1\ 2)$
$\sigma_5(1) = 2, \sigma_5(2) = 3, \sigma_5(3) = 1$	$(1\ 2\ 3)$
$\sigma_6(1) = 3, \sigma_6(2) = 1, \sigma_6(3) = 2$	$(1\ 3\ 2)$

For any $\sigma \in S_n$, the cycle decomposition of σ^{-1} is obtained by writing the numbers in each cycle of the cycle decomposition of σ in reverse order. For example, if $\sigma = (1\ 12\ 8\ 10\ 4)(2\ 13)(5\ 11\ 7)(6\ 9)$ is the element of S_{13} described before then

$$\sigma^{-1} = (4\ 10\ 8\ 12\ 1)(13\ 2)(7\ 11\ 5)(9\ 6).$$

Computing products in S_n is straightforward, keeping in mind that when computing $\sigma \circ \tau$ in S_n one reads the permutations from *right to left*. One simply "follows" the elements under the successive permutations. For example, in the product $(1\ 2\ 3) \circ (1\ 2)(3\ 4)$ the number 1 is sent to 2 by the first permutation, then 2 is sent to 3 by the second permutation, hence the composite maps 1 to 3. To compute the cycle decomposition of the product we need next to see what happens to 3. It is sent first to 4, then 4 is fixed, so 3 is mapped to 4 by the composite map. Similarly, 4 is first mapped to 3 then 3 is mapped to 1, completing this cycle in the product: $(1\ 3\ 4)$. Finally, 2 is sent to 1, then 1 is sent to 2 so 2 is fixed by this product and so $(1\ 2\ 3) \circ (1\ 2)(3\ 4) = (1\ 3\ 4)$ is the cycle decomposition of the product.

As additional examples,

$$(12) \circ (13) = (1\ 3\ 2) \qquad \text{and} \qquad (1\ 3) \circ (1\ 2) = (1\ 2\ 3).$$

In particular this shows that

$$S_n \textit{ is a non-abelian group for all } n \geq 3.$$

Each cycle

$$(a_1\ a_2 \ldots a_m)$$

in a cycle decomposition can be viewed as the permutation which cyclically permutes $a_1, a_2, \ldots, a_m$ and fixes all other integers. Since disjoint cycles permute numbers which lie in disjoint sets it follows that

$$\textit{disjoint cycles commute.}$$

Thus rearranging the cycles in any product of disjoint cycles (in particular, in a cycle decomposition) does not change the permutation.

Also, since a given cycle, $(a_1\ a_2 \ldots a_m)$, permutes $\{a_1, a_2, \ldots, a_m\}$ cyclically, the numbers in the cycle itself can be cyclically permuted without altering the permutation, i.e.

$$\begin{aligned}(a_1\ a_2 \ldots a_m) &= (a_2\ a_3 \ldots a_{m-1}\ a_1) = (a_3\ a_4 \ldots a_{m-1}\ a_1\ a_2) = \ldots \\ &= (a_m\ a_1\ a_2 \ldots a_{m-1}).\end{aligned}$$

Thus, for instance, $(1\ 2) = (2\ 1)$ and $(1\ 2\ 3\ 4) = (3\ 4\ 1\ 2)$. By convention, the smallest number appearing in the cycle is usually written first.

One must exercise some care working with cycles since a permutation may be written in many ways as an arbitrary product of cycles. For instance, in S_3, $(1\ 2\ 3) = (1\ 2)(2\ 3) = (1\ 3)(1\ 3\ 2)(1\ 3)$ etc. But, (as we shall prove) the cycle decomposition of each permutation is the *unique* way of expressing a permutation as a product of disjoint cycles (up to rearranging its cycles and cyclically permuting the numbers within each cycle). Reducing an arbitrary product of cycles to a product of disjoint cycles allows us to determine at a glance whether or not two permutations are the same. Another advantage to this notation is that it is an exercise (outlined below) to prove that *the order of a permutation is the l.c.m. of the lengths of the cycles in its cycle decomposition.*

EXERCISES

1. Let σ be the permutation

$$1 \mapsto 3 \qquad 2 \mapsto 4 \qquad 3 \mapsto 5 \qquad 4 \mapsto 2 \qquad 5 \mapsto 1$$

and let τ be the permutation

$$1 \mapsto 5 \qquad 2 \mapsto 3 \qquad 3 \mapsto 2 \qquad 4 \mapsto 4 \qquad 5 \mapsto 1.$$

Find the cycle decompositions of each of the following permutations: σ, τ, σ^2, $\sigma\tau$, $\tau\sigma$, and $\tau^2\sigma$.

2. Let σ be the permutation

$$\begin{array}{lllll} 1 \mapsto 13 & 2 \mapsto 2 & 3 \mapsto 15 & 4 \mapsto 14 & 5 \mapsto 10 \\ 6 \mapsto 6 & 7 \mapsto 12 & 8 \mapsto 3 & 9 \mapsto 4 & 10 \mapsto 1 \\ 11 \mapsto 7 & 12 \mapsto 9 & 13 \mapsto 5 & 14 \mapsto 11 & 15 \mapsto 8 \end{array}$$

and let τ be the permutation

$$\begin{array}{lllll} 1 \mapsto 14 & 2 \mapsto 9 & 3 \mapsto 10 & 4 \mapsto 2 & 5 \mapsto 12 \\ 6 \mapsto 6 & 7 \mapsto 5 & 8 \mapsto 11 & 9 \mapsto 15 & 10 \mapsto 3 \\ 11 \mapsto 8 & 12 \mapsto 7 & 13 \mapsto 4 & 14 \mapsto 1 & 15 \mapsto 13. \end{array}$$

Find the cycle decompositions of each of the following permutations: σ, τ, σ^2, $\sigma\tau$, $\tau\sigma$, and $\tau^2\sigma$.

3. For each of the permutations whose cycle decompositions were computed in the preceding two exercises compute its order.

4. Compute the order of each of the elements in the following groups:
(a) S_3 **(b)** S_4.

5. Find the order of $(1\ 12\ 8\ 10\ 4)(2\ 13)(5\ 11\ 7)(6\ 9)$.

6. Write out the cycle decomposition of each element of order 4 in S_4.

7. Write out the cycle decomposition of each element of order 2 in S_4.

8. Prove that if $\Omega = \{1, 2, 3, \ldots\}$ then S_Ω is an infinite group (do not say '$\infty! = \infty$').

9. **(a)** Let σ be the 12-cycle $(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12)$. For which positive integers i is σ^i also a 12-cycle?
(b) Let τ be the 8-cycle $(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$. For which positive integers i is τ^i also an 8-cycle?
(c) Let ω be the 14-cycle $(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14)$. For which positive integers i is ω^i also a 14-cycle?

10. Prove that if σ is the m-cycle $(a_1\ a_2\ \ldots\ a_m)$, then for all $i \in \{1, 2, \ldots, m\}$, $\sigma^i(a_k) = a_{k+i}$, where $k + i$ is replaced by its least residue mod m when $k + i > m$. Deduce that $|\sigma| = m$.

11. Let σ be the m-cycle $(1\ 2\ \ldots\ m)$. Show that σ^i is also an m-cycle if and only if i is relatively prime to m.

12. **(a)** If $\tau = (1\ 2)(3\ 4)(5\ 6)(7\ 8)(9\ 10)$ determine whether there is an n–cycle σ $(n \geq 10)$ with $\tau = \sigma^k$ for some integer k.
(b) If $\tau = (1\ 2)(3\ 4\ 5)$ determine whether there is an n–cycle σ $(n \geq 5)$ with $\tau = \sigma^k$ for some integer k.

13. Show that an element has order 2 in S_n if and only if its cycle decomposition is a product of commuting 2-cycles.

14. Let p be a prime. Show that an element has order p in S_n if and only if its cycle decomposition is a product of commuting p-cycles. Show by an explicit example that this need not be the case if p is not prime.

15. Prove that the order of an element in S_n equals the least common multiple of the lengths of the cycles in its cycle decomposition. [Use Exercise 10 and Exercise 21 of Section 1.]

16. Show that if $n \geq m$ then the number of m-cycles in S_n is given by

$$\frac{n(n-1)(n-2)\ldots(n-m+1)}{m}.$$

[Count the number of ways of forming an m-cycle and divide by the number of representations of a particular m-cycle.]

17. Show that if $n \geq 4$ then the number of permutations in S_n which are the product of two disjoint 2-cycles is

$$\frac{n(n-1)(n-2)(n-3)}{8}.$$

18. Find all numbers n such that S_5 contains an element of order n. [Use Exercise 15.]

19. Find all numbers n such that S_7 contains an element of order n. [Use Exercise 15.]

1.4 MATRIX GROUPS

In this section we introduce the notion of matrix groups where the coefficients come from fields. This example of a family of groups will be used for illustrative purposes in Part I and will be studied in more detail in the chapters on vector spaces.

A *field* is the "smallest" mathematical structure in which we can perform all the arithmetic operations $+$, $-$, $\times$, and $\div$ (division by nonzero elements), so in particular every nonzero element must have a multiplicative inverse. We shall study fields more thoroughly later and in this part of the text the only fields F we shall encounter will be $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{Z}/p\mathbb{Z}$, where p is a prime. The example $\mathbb{Z}/p\mathbb{Z}$ is a finite field, which, to emphasize that it is a field, we shall denote by $\mathbb{F}_p$. For the sake of completeness we include here the precise definition of a field.

Definition

(1) A *field* is a triple $(F, +, \cdot)$ such that $(F, +)$ is an abelian group (call its identity 0) and $(F - \{0\}, \cdot)$ is also an abelian group, and the following *distributive* law holds:

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c), \qquad \text{for all } a, b, c \in F.$$

(2) For any field F let $F^\times = F - \{0\}$.

All the vector space theory, the theory of matrices and linear transformations and the theory of determinants when the scalars come from $\mathbb{R}$ is true, *mutatis mutandis*, when the scalars come from an arbitrary field F. When we use this theory in Part I we shall state explicitly what facts on fields we are assuming.

For each $n \in \mathbb{Z}^+$ let $GL_n(F)$ be the set of all $n \times n$ matrices whose entries come from F and whose determinant is nonzero, i.e.

$$GL_n(F) = \{A \mid A \text{ is an } n \times n \text{ matrix with entries from } F \text{ and } \det(A) \neq 0\},$$

where the determinant of any matrix A with entries from F can be computed by the same formulas used when $F = \mathbb{R}$. For arbitrary $n \times n$ matrices A and B let AB be the product of these matrices as computed by the same rules as when $F = \mathbb{R}$. This product is associative. Also, since $\det(AB) = \det(A) \cdot \det(B)$, it follows that if $\det(A) \neq 0$ and $\det(B) \neq 0$, then $\det(AB) \neq 0$, so $GL_n(F)$ is closed under matrix multiplication. Furthermore, $\det(A) \neq 0$ if and only if A has a matrix inverse (and this inverse can be computed by the same adjoint formula used when $F = \mathbb{R}$), so each $A \in GL_n(F)$ has an inverse, A^{-1}, in $GL_n(F)$:

$$AA^{-1} = A^{-1}A = I,$$

where I is the $n \times n$ identity matrix. Thus $GL_n(F)$ is a group under matrix multiplication, called the *general linear group of degree* n.

The following results will be proved in Part III but are recorded now for convenience:

(1) if F is a field and $|F| < \infty$, then $|F| = p^m$ for some prime p and integer m

(2) if $|F| = q < \infty$, then $|GL_n(F)| = (q^n - 1)(q^n - q)(q^n - q^2)\dots(q^n - q^{n-1})$.

EXERCISES

Let F be a field and let $n \in \mathbb{Z}^+$.

1. Prove that $|GL_2(\mathbb{F}_2)| = 6$.
2. Write out all the elements of $GL_2(\mathbb{F}_2)$ and compute the order of each element.
3. Show that $GL_2(\mathbb{F}_2)$ is non-abelian.
4. Show that if n is not prime then $\mathbb{Z}/n\mathbb{Z}$ is not a field.
5. Show that $GL_n(F)$ is a finite group if and only if F has a finite number of elements.
6. If $|F| = q$ is finite prove that $|GL_n(F)| < q^{n^2}$.
7. Let p be a prime. Prove that the order of $GL_2(\mathbb{F}_p)$ is $p^4 - p^3 - p^2 + p$ (do not just quote the order formula in this section). [Subtract the number of 2×2 matrices which are *not* invertible from the total number of 2×2 matrices over $\mathbb{F}_p$. You may use the fact that a 2×2 matrix is not invertible if and only if one row is a multiple of the other.]
8. Show that $GL_n(F)$ is non-abelian for any $n \geq 2$ and any F.
9. Prove that the binary operation of matrix multiplication of 2×2 matrices with real number entries is associative.
10. Let $G = \{\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in \mathbb{R},\ a \neq 0,\ c \neq 0\}$.
    - **(a)** Compute the product of $\begin{pmatrix} a_1 & b_1 \\ 0 & c_1 \end{pmatrix}$ and $\begin{pmatrix} a_2 & b_2 \\ 0 & c_2 \end{pmatrix}$ to show that G is closed under matrix multiplication.
    - **(b)** Find the matrix inverse of $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$ and deduce that G is closed under inverses.
    - **(c)** Deduce that G is a subgroup of $GL_2(\mathbb{R})$ (cf. Exercise 26, Section 1).
    - **(d)** Prove that the set of elements of G whose two diagonal entries are equal (i.e. $a = c$) is also a subgroup of $GL_2(\mathbb{R})$.

1.5 THE QUATERNION GROUP

The *quaternion group*, Q_8, is defined by

$$Q_8 = \{1, -1, i, -i, j, -j, k, -k\}$$

with product $\cdot$ computed as follows:

$$1 \cdot a = a \cdot 1 = a, \qquad \text{for all } a \in Q_8$$

$$(-1) \cdot (-1) = 1, \qquad (-1) \cdot a = a \cdot (-1) = -a, \qquad \text{for all } a \in Q_8$$

$$i \cdot i = j \cdot j = k \cdot k = -1$$

$$i \cdot j = k, \qquad j \cdot i = -k$$

$$j \cdot k = i, \qquad k \cdot j = -i$$

$$k \cdot i = j, \qquad i \cdot k = -j.$$

As usual, we shall henceforth write ab for $a \cdot b$. It is tedious to check the associative law (we shall prove this later by less computational means), but the other axioms are easily checked. Note that Q_8 is a non-abelian group of order 8.

EXERCISES

1. Compute the order of each of the elements in Q_8.
2. Write out the group tables for S_3, D_8 and Q_8.
3. Find a set of generators for $\mathbb{Z}$.
4. Find a set of generators and relations for $\mathbb{Z}/n\mathbb{Z}$.
5. Find a set of generators and relations for Q_8.
6. Find a set of generators and relations for S_3.

1.6 HOMOMORPHISMS AND ISOMORPHISMS

In this section we make precise the notion of when two groups "look the same", that is, have exactly the same group-theoretic structure. This is the notion of an *isomorphism* between two groups. We first define the notion of a *homomorphism* about which we shall have a great deal more to say later.

Definition. Let $(G, \star)$ and $(H, \diamond)$ be groups. A map $\varphi : G \to H$ such that

$$\varphi(x \star y) = \varphi(x) \diamond \varphi(y), \qquad \text{for all } x, y \in G$$

is called a *homomorphism*.

When the group operations for G and H are not explicitly written, the homomorphism condition becomes simply

$$\varphi(xy) = \varphi(x)\varphi(y)$$

but it is important to keep in mind that the product xy on the left is computed in G and the product $\varphi(x)\varphi(y)$ on the right is computed in H. Intuitively, a map φ is a homomorphism if it respects the group structures of its domain and codomain.

Definition. The map $\varphi : G \to H$ is called an *isomorphism* and G and H are said to be *isomorphic* or of the same *isomorphism type*, written $G \cong H$, if

(1) φ is a homomorphism (i.e. $\varphi(xy) = \varphi(x)\varphi(y)$), and
(2) φ is a bijection.

In other words, the groups G and H are isomorphic if there is a bijection between them which preserves the group operations. Intuitively, G and H are the same group except that the elements and the operations may be written differently in G and H. Thus any property which G has which depends only on the group structure of G (i.e. can be derived from the group axioms — for example, commutativity of the group)

also holds in H. Note that this formally justifies writing all our group operations as $\cdot$ since changing the symbol of the operation does not change the isomorphism type.

Examples

(1) For any group G, $G \cong G$. The identity map provides an obvious isomorphism but not, in general, the *only* isomorphism from G to itself. More generally, let $\mathcal{G}$ be any nonempty collection of groups. It is easy to check that the relation $\cong$ is an equivalence relation on $\mathcal{G}$ and the equivalence classes are called *isomorphism classes*. This accounts for the somewhat symmetric wording of the definition of "isomorphism".

(2) The exponential map $\exp : \mathbb{R} \to \mathbb{R}^+$ defined by $\exp(x) = e^x$, where e is the base of the natural logarithm, is an isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}^+, \times)$. Exp is a bijection since it has an inverse function (namely $\log_e$) and exp preserves the group operations since $e^{x+y} = e^x e^y$. In this example both the elements and the operations are different yet the two groups are isomorphic, that is, as groups they have identical structures.

(3) In this example we show that the isomorphism type of a symmetric group depends only on the cardinality of the underlying set being permuted.

Let Δ and Ω be nonempty sets. The symmetric groups S_Δ and S_Ω are isomorphic if $|\Delta| = |\Omega|$. We can see this intuitively as follows: given that $|\Delta| = |\Omega|$, there is a bijection θ from Δ onto Ω. Think of the elements of Δ and Ω as being glued together via θ, i.e. each $x \in \Delta$ is glued to $\theta(x) \in \Omega$. To obtain a map $\varphi : S_\Delta \to S_\Omega$ let $\sigma \in S_\Delta$ be a permutation of Δ and let $\varphi(\sigma)$ be the permutation of Ω which moves the elements of Ω in the same way σ moves the corresponding glued elements of Δ; that is, if $\sigma(x) = y$, for some $x, y \in \Delta$, then $\varphi(\sigma)(\theta(x)) = \theta(y)$ in Ω . Since the set bijection θ has an inverse, one can easily check that the map between symmetric groups also has an inverse. The precise technical definition of the map φ and the straightforward, albeit tedious, checking of the properties which ensure φ is an isomorphism are relegated to the following exercises.

Conversely, if $S_\Delta \cong S_\Omega$, then $|\Delta| = |\Omega|$; we prove this only when the underlying sets are finite (when both Δ and Ω are infinite sets the proof is harder and will be given as an exercise in Chapter 4). Since any isomorphism between two groups G and H is, a priori, a bijection between them, a necessary condition for isomorphism is $|S_\Delta| = |S_\Omega|$. When Δ is a finite set of order n, then $|S_\Delta| = n!$. We actually only proved this for S_n, however the same reasoning applies for S_Δ. Similarly, if Ω is a finite set of order m, then $|S_\Omega| = m!$. Thus if S_Δ and S_Ω are isomorphic then $n! = m!$, so $m = n$, i.e. $|\Delta| = |\Omega|$.

Many more examples of isomorphisms will appear throughout the text. When we study different structures (rings, fields, vector spaces, etc.) we shall formulate corresponding notions of isomorphisms between respective structures. One of the central problems in mathematics is to determine what properties of a structure specify its isomorphism type (i.e. to prove that if G is an object with some structure (such as a group) and G has property $\mathcal{P}$, then any other similarly structured object (group) X with property $\mathcal{P}$ is isomorphic to G). Theorems of this type are referred to as *classification theorems*. For example, we shall prove that

$$\textit{any non-abelian group of order } 6 \textit{ is isomorphic to } S_3$$

(so here G is the group S_3 and $\mathcal{P}$ is the property "non-abelian and of order 6"). From this classification theorem we obtain $D_6 \cong S_3$ and $GL_2(\mathbb{F}_2) \cong S_3$ without having

to find explicit maps between these groups. Note that it is not true that any group of order 6 is isomorphic to S_3. In fact we shall prove that up to isomorphism there are precisely two groups of order 6: S_3 and $\mathbb{Z}/6\mathbb{Z}$ (i.e. any group of order 6 is isomorphic to one of these two groups and S_3 is not isomorphic to $\mathbb{Z}/6\mathbb{Z}$). Note that the conclusion is less specific (there are two possible types); however, the hypotheses are easier to check (namely, check to see if the order is 6). Results of the latter type are also referred to as classifications. Generally speaking it is subtle and difficult, even in specific instances, to determine whether or not two groups (or other mathematical objects) are isomorphic — constructing an explicit map between them which preserves the group operations or proving no such map exists is, except in tiny cases, computationally unfeasible as indicated already in trying to prove the above classification of groups of order 6 without further theory.

It is occasionally easy to see that two given groups are *not* isomorphic. For example, the exercises below assert that if $\varphi : G \to H$ is an isomorphism, then, in particular,

(a) $|G| = |H|$

(b) G is abelian if and only if H is abelian

(c) for all $x \in G$, $|x| = |\varphi(x)|$.

Thus S_3 and $\mathbb{Z}/6\mathbb{Z}$ are not isomorphic (as indicated above) since one is abelian and the other is not. Also, $(\mathbb{R} - \{0\}, \times)$ and $(\mathbb{R}, +)$ cannot be isomorphic because in $(\mathbb{R} - \{0\}, \times)$ the element -1 has order 2 whereas $(\mathbb{R}, +)$ has no element of order 2, contrary to (c).

Finally, we record one very useful fact that we shall prove later (when we discuss free groups) dealing with the question of homomorphisms and isomorphisms between two groups given by generators and relations:

Let G be a finite group of order n for which we have a presentation and let $S = \{s_1, \ldots, s_m\}$ be the generators. Let H be another group and $\{r_1, \ldots, r_m\}$ be elements of H. Suppose that any relation satisfied in G by the s_i is also satisfied in H when each s_i is replaced by r_i. Then there is a (unique) homomorphism $\varphi : G \to H$ which maps s_i to r_i. If we have a presentation for G, then we need only check the relations specified by this presentation (since, by definition of a presentation, every relation can be deduced from the relations given in the presentation). If H is generated by the elements $\{r_1, \ldots, r_m\}$, then φ is surjective (any product of the r_i's is the image of the corresponding product of the s_i's). If, in addition, H has the same (finite) order as G, then any surjective map is necessarily injective, i.e. φ is an isomorphism: $G \cong H$. Intuitively, we can map the generators of G to any elements of H and obtain a homomorphism provided that the relations in G are still satisfied.

Readers may already be familiar with the corresponding statement for vector spaces. Suppose V is a finite dimensional vector space of dimension n with basis S and W is another vector space. Then we can specify a linear transformation from V to W by mapping the elements of S to arbitrary vectors in W (here there are no relations to satisfy). If W is also of dimension n and the chosen vectors in W span W (and so are a basis for W) then this linear transformation is invertible (a vector space isomorphism).

Examples

(1) Recall that $D_{2n} = \langle r, s \mid r^n = s^2 = 1, sr = r^{-1}s \rangle$. Suppose H is a group containing elements a and b with $a^n = 1, b^2 = 1$ and $ba = a^{-1}b$. Then there is a homomorphism from D_{2n} to H mapping r to a and s to b. For instance, let k be an integer dividing n with $k \geq 3$ and let $D_{2k} = \langle r_1, s_1 \mid r_1^k = s_1^2 = 1, s_1 r_1 = r_1^{-1} s_1 \rangle$. Define

$$\varphi : D_{2n} \to D_{2k} \quad \text{by} \quad \varphi(r) = r_1 \text{ and } \varphi(s) = s_1.$$

If we write $n = km$, then since $r_1^k = 1$, also $r_1^n = (r_1^k)^m = 1$. Thus the three relations satisfied by r, s in D_{2n} are satisfied by r_1, s_1 in D_{2k}. Thus φ extends (uniquely) to a homomorphism from D_{2n} to D_{2k}. Since $\{r_1, s_1\}$ generates D_{2k}, φ is surjective. This homomorphism is not an isomorphism if $k < n$.

(2) Following up on the preceding example, let $G = D_6$ be as presented above. Check that in $H = S_3$ the elements $a = (1\,2\,3)$ and $b = (1\,2)$ satisfy the relations: $a^3 = 1$, $b^2 = 1$ and $ba = ab^{-1}$. Thus there is a homomorphism from D_6 to S_3 which sends $r \mapsto a$ and $s \mapsto b$. One may further check that S_3 is generated by a and b, so this homomorphism is surjective. Since D_6 and S_3 both have order 6, this homomorphism is an isomorphism: $D_6 \cong S_3$.

Note that the element a in the examples above need not have *order* n (i.e. n need not be the *smallest* power of a giving the identity in H) and similarly b need not have order 2 (for example b could well be the identity if $a = a^{-1}$). This allows us to more easily construct homomorphisms and is in keeping with the idea that the generators and relations for a group G constitute a complete set of data for the group structure of G.

EXERCISES

Let G and H be groups.

1. Let $\varphi : G \to H$ be a homomorphism.
   (a) Prove that $\varphi(x^n) = \varphi(x)^n$ for all $n \in \mathbb{Z}^+$.
   (b) Do part (a) for $n = -1$ and deduce that $\varphi(x^n) = \varphi(x)^n$ for all $n \in \mathbb{Z}$.

2. If $\varphi : G \to H$ is an isomorphism, prove that $|\varphi(x)| = |x|$ for all $x \in G$. Deduce that any two isomorphic groups have the same number of elements of order n for each $n \in \mathbb{Z}^+$. Is the result true if φ is only assumed to be a homomorphism?

3. If $\varphi : G \to H$ is an isomorphism, prove that G is abelian if and only if H is abelian. If $\varphi : G \to H$ is a homomorphism, what additional conditions on φ (if any) are sufficient to ensure that if G is abelian, then so is H?

4. Prove that the multiplicative groups $\mathbb{R} - \{0\}$ and $\mathbb{C} - \{0\}$ are not isomorphic.

5. Prove that the additive groups $\mathbb{R}$ and $\mathbb{Q}$ are not isomorphic.

6. Prove that the additive groups $\mathbb{Z}$ and $\mathbb{Q}$ are not isomorphic.

7. Prove that D_8 and Q_8 are not isomorphic.

8. Prove that if $n \neq m$, S_n and S_m are not isomorphic.

9. Prove that D_{24} and S_4 are not isomorphic.

10. Fill in the details of the proof that the symmetric groups S_Δ and S_Ω are isomorphic if $|\Delta| = |\Omega|$ as follows: let $\theta : \Delta \to \Omega$ be a bijection. Define

$$\varphi : S_\Delta \to S_\Omega \qquad \text{by} \qquad \varphi(\sigma) = \theta \circ \sigma \circ \theta^{-1} \quad \text{for all } \sigma \in S_\Delta$$

and prove the following:

(a) φ is well defined, that is, if σ is a permutation of Δ then $\theta \circ \varphi \circ \theta^{-1}$ is a permutation of Ω.

(b) φ is a bijection from S_Δ onto S_Ω. [Find a 2-sided inverse for φ.]

(c) φ is a homomorphism, that is, $\varphi(\sigma \circ \tau) = \varphi(\sigma) \circ \varphi(\tau)$.

Note the similarity to the *change of basis* or *similarity* transformations for matrices (we shall see the connections between these later in the text).

11. Let A and B be groups. Prove that $A \times B \cong B \times A$.

12. Let A, B, and C be groups and let $G = A \times B$ and $H = B \times C$. Prove that $G \times C \cong A \times H$.

13. Let G and H be groups and let $\varphi : G \to H$ be a homomorphism. Prove that the image of φ, $\varphi(G)$, is a subgroup of H (cf. Exercise 26 of Section 1). Prove that if φ is injective then $G \cong \varphi(G)$.

14. Let G and H be groups and let $\varphi : G \to H$ be a homomorphism. Define the *kernel* of φ to be $\{g \in G \mid \varphi(g) = 1_H\}$ (so the kernel is the set of elements in G which map to the identity of H, i.e. is the fiber over the identity of H). Prove that the kernel of φ is a subgroup (cf. Exercise 26 of Section 1) of G. Prove that φ is injective if and only if the kernel of φ is the identity subgroup of G.

15. Define a map $\pi : \mathbb{R}^2 \to \mathbb{R}$ by $\pi((x, y)) = x$. Prove that π is a homomorphism and find the kernel of π (cf. Exercise 14).

16. Let A and B be groups and let G be their direct product, $A \times B$. Prove that the maps $\pi_1 : G \to A$ and $\pi_2 : G \to B$ defined by $\pi_1((a, b)) = a$ and $\pi_2((a, b)) = b$ are homomorphisms and find their kernels (cf. Exercise 14).

17. Let G be any group. Prove that the map from G to itself defined by $g \mapsto g^{-1}$ is a homomorphism if and only if G is abelian.

18. Let G be any group. Prove that the map from G to itself defined by $g \mapsto g^2$ is a homomorphism if and only if G is abelian.

19. Let $G = \{z \in \mathbb{C} \mid z^n = 1 \text{ for some } n \in \mathbb{Z}^+\}$. Prove that for any fixed integer $k > 1$ the map from G to itself defined by $z \mapsto z^k$ is a surjective homomorphism but is not an isomorphism.

20. Let G be a group and let $\mathrm{Aut}(G)$ be the set of all isomorphisms from G onto G. Prove that $\mathrm{Aut}(G)$ is a group under function composition (called the *automorphism group* of G and the elements of $\mathrm{Aut}(G)$ are called *automorphisms* of G).

21. Prove that for each fixed nonzero $k \in \mathbb{Q}$ the map from $\mathbb{Q}$ to itself defined by $q \mapsto kq$ is an automorphism of $\mathbb{Q}$ (cf. Exercise 20).

22. Let A be an abelian group and fix some $k \in \mathbb{Z}$. Prove that the map $a \mapsto a^k$ is a homomorphism from A to itself. If $k = -1$ prove that this homomorphism is an isomorphism (i.e. is an automorphism of A).

23. Let G be a finite group which possesses an automorphism σ (cf. Exercise 20) such that $\sigma(g) = g$ if and only if $g = 1$. If σ^2 is the identity map from G to G, prove that G is abelian (such an automorphism σ is called *fixed point free* of order 2). [Show that every element of G can be written in the form $x^{-1}\sigma(x)$ and apply σ to such an expression.]

24. Let G be a finite group and let x and y be distinct elements of order 2 in G that generate G. Prove that $G \cong D_{2n}$, where $n = |xy|$. [See Exercise 6 in Section 2.]

25. Let $n \in \mathbb{Z}^+$, let r and s be the usual generators of D_{2n} and let $\theta = 2\pi/n$.

(a) Prove that the matrix $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ is the matrix of the linear transformation

which rotates the x, y plane about the origin in a counterclockwise direction by θ radians.

(b) Prove that the map $\varphi : D_{2n} \to GL_2(\mathbb{R})$ defined on generators by

$$\varphi(r) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \qquad \text{and} \qquad \varphi(s) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

extends to a homomorphism of D_{2n} into $GL_2(\mathbb{R})$.

(c) Prove that the homomorphism φ in part (b) is injective.

26. Let i and j be the generators of Q_8 described in Section 5. Prove that the map $\varphi : Q_8 \to GL_2(\mathbb{C})$ defined on generators by

$$\varphi(i) = \begin{pmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{pmatrix} \qquad \text{and} \qquad \varphi(j) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

extends to a homomorphism. Prove that φ is injective.

1.7 GROUP ACTIONS

In this section we introduce the precise definition of a group acting on a set and present some examples. Group actions will be a powerful tool which we shall use both for proving theorems for abstract groups and for unravelling the structure of specific examples. Moreover, the concept of an "action" is a theme which will recur throughout the text as a method for studying an algebraic object by seeing how it can act on other structures.

Definition. A *group action* of a group G on a set A is a map from $G \times A$ to A (written as $g \cdot a$, for all $g \in G$ and $a \in A$) satisfying the following properties:

(1) $g_1 \cdot (g_2 \cdot a) = (g_1 g_2) \cdot a$, for all $g_1, g_2 \in G$, $a \in A$, and

(2) $1 \cdot a = a$, for all $a \in A$.

We shall immediately become less formal and say G is a group acting on a set A. The expression $g \cdot a$ will usually be written simply as ga when there is no danger of confusing this map with, say, the group operation (remember, $\cdot$ is not a binary operation and ga is always a member of A). Note that on the left hand side of the equation in property (1) $g_2 \cdot a$ is an element of A so it makes sense to act on this by g_1. On the right hand side of this equation the product $(g_1 g_2)$ is taken in G and the resulting group element acts on the set element a.

Before giving some examples of group actions we make some observations. Let the group G act on the set A. For each fixed $g \in G$ we get a map σ_g defined by

$$\sigma_g : A \to A$$
$$\sigma_g(a) = g \cdot a.$$

We prove two important facts:

(i) for each fixed $g \in G$, σ_g is a *permutation* of A, and

(ii) the map from G to S_A defined by $g \mapsto \sigma_g$ is a homomorphism.

To see that σ_g is a permutation of A we show that as a set map from A to A it has a 2-sided inverse, namely $\sigma_{g^{-1}}$ (it is then a permutation by Proposition 1 of Section 0.1). For all $a \in A$

$$\begin{aligned}
(\sigma_{g^{-1}} \circ \sigma_g)(a) &= \sigma_{g^{-1}}(\sigma_g(a)) && \text{(by definition of function composition)}\\
&= g^{-1} \cdot (g \cdot a) && \text{(by definition of } \sigma_{g^{-1}} \text{ and } \sigma_g)\\
&= (g^{-1}g) \cdot a && \text{(by property (1) of an action)}\\
&= 1 \cdot a = a && \text{(by property (2) of an action).}
\end{aligned}$$

This proves $\sigma_{g^{-1}} \circ \sigma_g$ is the identity map from A to A. Since g was arbitrary, we may interchange the roles of g and g^{-1} to obtain $\sigma_g \circ \sigma_{g^{-1}}$ is also the identity map on A. Thus σ_g has a 2-sided inverse, hence is a permutation of A.

To check assertion (ii) above let $\varphi : G \to S_A$ be defined by $\varphi(g) = \sigma_g$. Note that part (i) shows that σ_g is indeed an element of S_A. To see that φ is a homomorphism we must prove $\varphi(g_1g_2) = \varphi(g_1) \circ \varphi(g_2)$ (recall that S_A is a group under function composition). The permutations $\varphi(g_1g_2)$ and $\varphi(g_1) \circ \varphi(g_2)$ are equal if and only if their values agree on every element $a \in A$. For all $a \in A$

$$\begin{aligned}
\varphi(g_1g_2)(a) &= \sigma_{g_1g_2}(a) && \text{(by definition of } \varphi)\\
&= (g_1g_2) \cdot a && \text{(by definition of } \sigma_{g_1g_2})\\
&= g_1 \cdot (g_2 \cdot a) && \text{(by property (1) of an action)}\\
&= \sigma_{g_1}(\sigma_{g_2}(a)) && \text{(by definition of } \sigma_{g_1} \text{ and } \sigma_{g_2})\\
&= (\varphi(g_1) \circ \varphi(g_2))(a) && \text{(by definition of } \varphi).
\end{aligned}$$

This proves assertion (ii) above.

Intuitively, a group action of G on a set A just means that every element g in G acts as a permutation on A in a manner consistent with the group operations in G; assertions (i) and (ii) above make this precise. The homomorphism from G to S_A given above is called the *permutation representation* associated to the given action. It is easy to see that this process is reversible in the sense that: if $\varphi : G \to S_A$ is any homomorphism from a group G to the symmetric group on a set A, then the map from $G \times A$ to A defined by

$$g \cdot a = \varphi(g)(a) \qquad \text{for all } g \in G, \text{ and all } a \in A$$

satisfies the properties of a group action of G on A. Thus actions of a group G on a set A and the homomorphisms from G into the symmetric group S_A are in bijective correspondence (i.e. are essentially the same notion, phrased in different terminology).

We should also note that the definition of an action might have been more precisely named a *left* action since the group elements appear on the left of the set elements. We could similarly define the notion of a *right* action.

Examples

Let G be a group and A a nonempty set. In each of the following examples the check of properties (1) and (2) of an action are left as exercises.

(1) Let $ga = a$, for all $g \in G$, $a \in A$. Properties (1) and (2) of a group action follow immediately. This action is called the *trivial action* and G is said to *act trivially* on A. Note that *distinct* elements of G induce the *same* permutation on A (in this case the

identity permutation). The associated permutation representation $G \to S_A$ is the trivial homomorphism which maps every element of G to the identity.

If G acts on a set B and distinct elements of G induce *distinct* permutations of B, the action is said to be *faithful*. A faithful action is therefore one in which the associated permutation representation is injective.

The *kernel* of the action of G on B is defined to be $\{g \in G \mid gb = b \text{ for all } b \in B\}$, namely the elements of G which fix *all* the elements of B. For the trivial action, the kernel of the action is all of G and this action is not faithful when $|G| > 1$.

(2) The axioms for a vector space V over a field F include the two axioms that the multiplicative group $F^\times$ act on the set V. Thus vector spaces are familiar examples of actions of multiplicative groups of fields where there is even more structure (in particular, V must be an abelian group) which can be exploited. In the special case when $V = \mathbb{R}^n$ and $F = \mathbb{R}$ the action is specified by

$$\alpha(r_1, r_2, \ldots, r_n) = (\alpha r_1, \alpha r_2, \ldots, \alpha r_n)$$

for all $\alpha \in \mathbb{R}$, $(r_1, r_2, \ldots, r_n) \in \mathbb{R}^n$, where αr_i is just multiplication of two real numbers.

(3) For any nonempty set A the symmetric group S_A acts on A by $\sigma \cdot a = \sigma(a)$, for all $\sigma \in S_A$, $a \in A$. The associated permutation representation is the identity map from S_A to itself.

(4) If we fix a labelling of the vertices of a regular n-gon, each element α of D_{2n} gives rise to a permutation σ_α of $\{1, 2, \ldots, n\}$ by the way the symmetry α permutes the corresponding vertices. The map of $D_{2n} \times \{1, 2, \ldots, n\}$ onto $\{1, 2, \ldots, n\}$ defined by $(\alpha, i) \to \sigma_\alpha(i)$ defines a group action of D_{2n} on $\{1, 2, \ldots, n\}$. In keeping with our notation for group actions we can now dispense with the formal and cumbersome notation $\sigma_\alpha(i)$ and write αi in its place. Note that this action is faithful: distinct symmetries of a regular n-gon induce distinct permutations of the vertices.

When $n = 3$ the action of D_6 on the three (labelled) vertices of a triangle gives an injective homomorphism from D_6 to S_3. Since these groups have the same order, this map must also be surjective, i.e. is an isomorphism: $D_6 \cong S_3$. This is another proof of the same fact we established via generators and relations in the preceding section. Geometrically it says that any permutation of the vertices of a triangle is a symmetry. The analogous statement is not true for any n-gon with $n \geq 4$ (just by order considerations we cannot have D_{2n} isomorphic to S_n for any $n \geq 4$).

(5) Let G be any group and let $A = G$. Define a map from $G \times A$ to A by $g \cdot a = ga$, for each $g \in G$ and $a \in A$, where ga on the right hand side is the product of g and a in the group G. This gives a group action of G on itself, where each (fixed) $g \in G$ permutes the elements of G by *left multiplication* :

$$g : a \mapsto ga \qquad \text{for all } a \in G$$

(or, if G is written additively, we get $a \mapsto g + a$ and call this *left translation*). This action is called the *left regular action* of G on itself. By the cancellation laws, this action is faithful (check this).

Other examples of actions are given in the exercises.

EXERCISES

1. Let F be a field. Show that the multiplicative group of nonzero elements of F (denoted by $F^\times$) acts on the set F by $g \cdot a = ga$, where $g \in F^\times$, $a \in F$ and ga is the usual product in F of the two field elements (state clearly which axioms in the definition of a field are used).

2. Show that the additive group $\mathbb{Z}$ acts on itself by $z \cdot a = z + a$ for all $z, a \in \mathbb{Z}$.
3. Show that the additive group $\mathbb{R}$ acts on the x, y plane $\mathbb{R} \times \mathbb{R}$ by $r \cdot (x, y) = (x + ry, y)$.
4. Let G be a group acting on a set A and fix some $a \in A$. Show that the following sets are subgroups of G (cf. Exercise 26 of Section 1):
   (a) the kernel of the action,
   (b) $\{g \in G \mid ga = a\}$ – this subgroup is called the *stabilizer* of a in G.
5. Prove that the kernel of an action of the group G on the set A is the same as the kernel of the corresponding permutation representation $G \to S_A$ (cf. Exercise 14 in Section 6).
6. Prove that a group G acts faithfully on a set A if and only if the kernel of the action is the set consisting only of the identity.
7. Prove that in Example 2 in this section the action is faithful.
8. Let A be a nonempty set and let k be a positive integer with $k \leq |A|$. The symmetric group S_A acts on the the set B consisting of all subsets of A of cardinality k by $\sigma \cdot \{a_1, \ldots, a_k\} = \{\sigma(a_1), \ldots, \sigma(a_k)\}$.
   (a) Prove that this is a group action.
   (b) Describe explicitly how the elements $(1\ 2)$ and $(1\ 2\ 3)$ act on the six 2-element subsets of $\{1, 2, 3, 4\}$.
9. Do both parts of the preceding exercise with "ordered k-tuples" in place of "k-element subsets", where the action on k-tuples is defined as above but with set braces replaced by parentheses (note that, for example, the 2-tuples (1,2) and (2,1) are different even though the sets $\{1, 2\}$ and $\{2, 1\}$ are the same, so the sets being acted upon are different).
10. With reference to the preceding two exercises determine:
    (a) for which values of k the action of S_n on k-element subsets is faithful, and
    (b) for which values of k the action of S_n on ordered k-tuples is faithful.
11. Write out the cycle decomposition of the eight permutations in S_4 corresponding to the elements of D_8 given by the action of D_8 on the vertices of a square (where the vertices of the square are labelled as in Section 2).
12. Assume n is an even positive integer and show that D_{2n} acts on the set consisting of pairs of opposite vertices of a regular n-gon. Find the kernel of this action (label vertices as usual).
13. Find the kernel of the left regular action.
14. Let G be a group and let $A = G$. Show that if G is non-abelian then the maps defined by $g \cdot a = ag$ for all $g, a \in G$ do *not* satisfy the axioms of a (left) group action of G on itself.
15. Let G be any group and let $A = G$. Show that the maps defined by $g \cdot a = ag^{-1}$ for all $g, a \in G$ *do* satisfy the axioms of a (left) group action of G on itself.
16. Let G be any group and let $A = G$. Show that the maps defined by $g \cdot a = gag^{-1}$ for all $g, a \in G$ *do* satisfy the axioms of a (left) group action (this action of G on itself is called *conjugation*).
17. Let G be a group and let G act on itself by left conjugation, so each $g \in G$ maps G to G by
$$x \mapsto gxg^{-1}.$$
    For fixed $g \in G$, prove that conjugation by g is an isomorphism from G onto itself (i.e. is an automorphism of G — cf. Exercise 20, Section 6). Deduce that x and gxg^{-1} have the same order for all x in G and that for any subset A of G, $|A| = |gAg^{-1}|$ (here $gAg^{-1} = \{gag^{-1} \mid a \in A\}$).
18. Let H be a group acting on a set A. Prove that the relation $\sim$ on A defined by
$$a \sim b \quad \text{if and only if} \quad a = hb \quad \text{for some } h \in H$$

is an equivalence relation. (For each $x \in A$ the equivalence class of x under $\sim$ is called the *orbit* of x under the action of H. The orbits under the action of H partition the set A.)

19. Let H be a subgroup (cf. Exercise 26 of Section 1) of the finite group G and let H act on G (here $A = G$) by left multiplication. Let $x \in G$ and let $\mathcal{O}$ be the orbit of x under the action of H. Prove that the map

$$H \to \mathcal{O} \qquad \text{defined by} \qquad h \mapsto hx$$

is a bijection (hence all orbits have cardinality $|H|$). From this and the preceding exercise deduce *Lagrange's Theorem*:

$$\textit{if } G \textit{ is a finite group and } H \textit{ is a subgroup of } G \textit{ then } |H| \textit{ divides } |G|.$$

20. Show that the group of rigid motions of a tetrahedron is isomorphic to a subgroup (cf. Exercise 26 of Section 1) of S_4.

21. Show that the group of rigid motions of a cube is isomorphic to S_4. [This group acts on the set of four pairs of opposite vertices.]

22. Show that the group of rigid motions of an octahedron is isomorphic to a subgroup (cf. Exercise 26 of Section 1) of S_4. [This group acts on the set of four pairs of opposite faces.] Deduce that the groups of rigid motions of a cube and an octahedron are isomorphic. (These groups are isomorphic because these solids are "dual" — see *Introduction to Geometry* by H. Coxeter, Wiley, 1961. We shall see later that the groups of rigid motions of the dodecahedron and icosahedron are isomorphic as well — these solids are also dual.)

23. Explain why the action of the group of rigid motions of a cube on the set of three pairs of opposite faces is not faithful. Find the kernel of this action.

2

Subgroups

2.1 DEFINITION AND EXAMPLES

One basic method for unravelling the structure of any mathematical object which is defined by a set of axioms is to study *subsets* of that object which also *satisfy the same axioms*. We begin this program by discussing subgroups of a group. A second basic method for unravelling structure is to study quotients of an object; the notion of a quotient group, which is a way (roughly speaking) of collapsing one group onto a smaller group, will be dealt with in the next chapter. Both of these themes will recur throughout the text as we study subgroups and quotient groups of a group, subrings and quotient rings of a ring, subspaces and quotient spaces of a vector space, etc.

Definition. Let G be a group. The subset H of G is a *subgroup* of G if H is nonempty and H is closed under products and inverses (i.e. $x, y \in H$ implies $x^{-1} \in H$ and $xy \in H$). If H is a subgroup of G we shall write $H \leq G$.

Subgroups of G are just subsets of G which are themselves groups with respect to the operation defined in G, i.e. the binary operation on G restricts to give a binary operation on H which is associative, has an identity in H, and has inverses in H for all the elements of H.

When we say that H is a subgroup of G we shall always mean that the operation for the group H is the operation on G restricted to H (in general it is possible that the subset H has the structure of a group with respect to some operation other than the operation on G restricted to H, cf. Example 5(a) following). As we have been doing for functions restricted to a subset, we shall denote the operation for G and the operation for the subgroup H by the same symbol. If $H \leq G$ and $H \neq G$ we shall write $H < G$ to emphasize that the containment is proper.

If H is a subgroup of G then, since the operation for H is the operation for G restricted to H, any equation in the subgroup H may also be viewed as an equation in the group G. Thus the cancellation laws for G imply that the identity for H is the same as the identity of G (in particular, every subgroup must contain 1, the identity of G) and the inverse of an element x in H is the same as the inverse of x when considered as an element of G (so the notation x^{-1} is unambiguous).

Examples

(1) $\mathbb{Z} \leq \mathbb{Q}$ and $\mathbb{Q} \leq \mathbb{R}$ with the operation of addition.

(2) Any group G has two subgroups: $H = G$ and $H = \{1\}$; the latter is called the *trivial subgroup* and will henceforth be denoted by 1.

(3) If $G = D_{2n}$ is the dihedral group of order $2n$, let H be $\{1, r, r^2, \ldots, r^{n-1}\}$, the set of all rotations in G. Since the product of two rotations is again a rotation and the inverse of a rotation is also a rotation it follows that H is a subgroup of D_{2n} of order n.

(4) The set of even integers is a subgroup of the group of all integers under addition.

(5) Some examples of subsets which are *not* subgroups:

(a) $\mathbb{Q} - \{0\}$ under multiplication is not a subgroup of $\mathbb{R}$ under addition even though both are groups and $\mathbb{Q} - \{0\}$ is a subset of $\mathbb{R}$; the operation of multiplication on $\mathbb{Q} - \{0\}$ is not the restriction of the operation of addition on $\mathbb{R}$.

(b) $\mathbb{Z}^+$ (under addition) is not a subgroup of $\mathbb{Z}$ (under addition) because although $\mathbb{Z}^+$ is closed under $+$, it does not contain the identity, 0, of $\mathbb{Z}$ and although each $x \in \mathbb{Z}^+$ has an additive inverse, $-x$, in $\mathbb{Z}$, $-x \notin \mathbb{Z}^+$, i.e. $\mathbb{Z}^+$ is not closed under the operation of taking inverses (in particular, $\mathbb{Z}^+$ is not a group under addition). For analogous reasons, $(\mathbb{Z} - \{0\}, \times)$ is not a subgroup of $(\mathbb{Q} - \{0\}, \times)$.

(c) D_6 is not a subgroup of D_8 since the former is not even a subset of the latter.

(6) The relation "is a subgroup of" is transitive: if H is a subgroup of a group G and K is a subgroup of H, then K is also a subgroup of G.

As we saw in Chapter 1, even for easy examples checking that all the group axioms (especially the associative law) hold for any given binary operation can be tedious at best. Once we know that we have a group, however, checking that a subset of it is (or is not) a subgroup is a much easier task, since all we need to check is closure under multiplication and under taking inverses. The next proposition shows that these can be amalgamated into a single test and also shows that for *finite* groups it suffices to check for closure under multiplication.

Proposition 1. (The Subgroup Criterion) A subset H of a group G is a subgroup if and only if

(1) $H \neq \emptyset$, and

(2) for all $x, y \in H$, $xy^{-1} \in H$.

Furthermore, if H is finite, then it suffices to check that H is nonempty and closed under multiplication.

Proof: If H is a subgroup of G, then certainly (1) and (2) hold because H contains the identity of G and the inverse of each of its elements and because H is closed under multiplication.

It remains to show conversely that if H satisfies both (1) and (2), then $H \leq G$. Let x be any element in H (such x exists by property (1)). Let $y = x$ and apply property (2) to deduce that $1 = xx^{-1} \in H$, so H contains the identity of G. Then, again by (2), since H contains 1 and x, H contains the element $1x^{-1}$, i.e. $x^{-1} \in H$ and H is closed under taking inverses. Finally, if x and y are any two elements of H, then H contains x and y^{-1} by what we have just proved, so by (2), H also contains

$x(y^{-1})^{-1} = xy$. Hence H is also closed under multiplication, which proves H is a subgroup of G.

Suppose now that H is finite and closed under multiplication and let x be any element in H. Then there are only finitely many distinct elements among $x, x^2, x^3, \ldots$ and so $x^a = x^b$ for some integers a, b with $b > a$. If $n = b - a$, then $x^n = 1$ so in particular every element $x \in H$ is of finite order. Then $x^{n-1} = x^{-1}$ is an element of H, so H is automatically also closed under inverses.

EXERCISES

Let G be a group.

1. In each of (a) to (e) prove that the specified subset is a subgroup of the given group:
   (a) the set of complex numbers of the form $a + ai$, $a \in \mathbb{R}$ (under addition)
   (b) the set of complex numbers of absolute value 1, i.e. the unit circle in the complex plane (under multiplication)
   (c) for fixed $n \in \mathbb{Z}^+$ the set of rational numbers whose denominators divide n (under addition)
   (d) for fixed $n \in \mathbb{Z}^+$ the set of rational numbers whose denominators are relatively prime to n (under addition)
   (e) the set of nonzero real numbers whose square is a rational number (under multiplication).
2. In each of (a) to (e) prove that the specified subset is *not* a subgroup of the given group:
   (a) the set of 2-cycles in S_n for $n \geq 3$
   (b) the set of reflections in D_{2n} for $n \geq 3$
   (c) for n a composite integer > 1 and G a group containing an element of order n, the set $\{x \in G \mid |x| = n\} \cup \{1\}$
   (d) the set of (positive and negative) odd integers in $\mathbb{Z}$ together with 0
   (e) the set of real numbers whose square is a rational number (under addition).
3. Show that the following subsets of the dihedral group D_8 are actually subgroups: (a) $\{1, r^2, s, sr^2\}$, (b) $\{1, r^2, sr, sr^3\}$.
4. Give an explicit example of a group G and an infinite subset H of G that is closed under the group operation but is not a subgroup of G.
5. Prove that G cannot have a subgroup H with $|H| = n - 1$, where $n = |G| > 2$.
6. Let G be an abelian group. Prove that $\{g \in G \mid |g| < \infty\}$ is a subgroup of G (called the *torsion subgroup* of G). Give an explicit example where this set is not a subgroup when G is non-abelian.
7. Fix some $n \in \mathbb{Z}$ with $n > 1$. Find the torsion subgroup (cf. the previous exercise) of $\mathbb{Z} \times (\mathbb{Z}/n\mathbb{Z})$. Show that the set of elements of infinite order together with the identity is *not* a subgroup of this direct product.
8. Let H and K be subgroups of G. Prove that $H \cup K$ is a subgroup if and only if either $H \subseteq K$ or $K \subseteq H$.
9. Let $G = GL_n(F)$, where F is any field. Define

$$SL_n(F) = \{A \in GL_n(F) \mid \det(A) = 1\}$$

(called the *special linear group*). Prove that $SL_n(F) \leq GL_n(F)$.

10. (a) Prove that if H and K are subgroups of G then so is their intersection $H \cap K$.
    (b) Prove that the intersection of an arbitrary nonempty collection of subgroups of G is again a subgroup of G (do not assume the collection is countable).
11. Let A and B be groups. Prove that the following sets are subgroups of the direct product $A \times B$:
    (a) $\{(a, 1) \mid a \in A\}$
    (b) $\{(1, b) \mid b \in B\}$
    (c) $\{(a, a) \mid a \in A\}$, where here we assume $B = A$ (called the *diagonal subgroup*).
12. Let A be an abelian group and fix some $n \in \mathbb{Z}$. Prove that the following sets are subgroups of A:
    (a) $\{a^n \mid a \in A\}$
    (b) $\{a \in A \mid a^n = 1\}$.
13. Let n be a fixed nonzero integer. Prove that $\{q \in \mathbb{Q} \mid nq = 2\}$ is not a subgroup of the additive group of rational numbers.
14. Show that $\{x \in D_{2n} \mid x^2 = 1\}$ is not a subgroup of D_{2n} (here $n \geq 3$).
15. Let $H_1 \leq H_2 \leq \cdots$ be an ascending chain of subgroups of G. Prove that $\cup_{i=1}^{\infty} H_i$ is a subgroup of G.
16. Let $n \in \mathbb{Z}^+$ and let F be a field. Prove that the set $\{(a_{ij}) \in GL_n(F) \mid a_{ij} = 0$ for all $i > j\}$ is a subgroup of $GL_n(F)$ (called the group of *upper triangular* matrices).
17. Let $n \in \mathbb{Z}^+$ and let F be a field. Prove that the set $\{(a_{ij}) \in GL_n(F) \mid a_{ij} = 0$ for all $i > j$, and $a_{ii} = 1$ for all $i\}$ is a subgroup of $GL_n(F)$.

2.2 CENTRALIZERS AND NORMALIZERS, STABILIZERS AND KERNELS

We now introduce some important families of subgroups of an arbitrary group G which in particular provide many examples of subgroups. Let A be any nonempty subset of G.

Centralizers: Define $C_G(A) = \{g \in G \mid gag^{-1} = a$ for all $a \in A\}$. This subset of G is called the *centralizer* of A in G. Since $gag^{-1} = a$ if and only if $ga = ag$, $C_G(A)$ is the set of elements of G which commute with every element of A.

We show $C_G(A)$ is a subgroup of G. First of all, $C_G(A) \neq \emptyset$ because $1 \in C_G(A)$: the definition of the identity specifies that $1a = a1$, for all $a \in G$ (in particular, for all $a \in A$) so 1 satisfies the defining condition for membership in $C_G(A)$. Secondly, assume $x, y \in C_G(A)$, that is, for all $a \in A$, $xax^{-1} = a$ and $yay^{-1} = a$ (note that this does *not* mean $xy = yx$). Observe first that since $yay^{-1} = a$, multiplying both sides of this first on the left by y^{-1}, then on the right by y and then simplifying gives $a = y^{-1}ay$, i.e. $y^{-1} \in C_G(A)$ so that $C_G(A)$ is closed under taking inverses. Now

$$\begin{aligned}
(xy)a(xy)^{-1} &= (xy)a(y^{-1}x^{-1}) && \text{(by Proposition 1.1(4) applied to } (xy)^{-1}\text{)}\\
&= x(yay^{-1})x^{-1} && \text{(by the associative law)}\\
&= xax^{-1} && \text{(since } y \in C_G(A)\text{)}\\
&= a && \text{(since } x \in C_G(A)\text{)}
\end{aligned}$$

so $xy \in C_G(A)$ and $C_G(A)$ is closed under products, hence $C_G(A) \leq G$.

In the special case when $A = \{a\}$ we shall write simply $C_G(a)$ instead of $C_G(\{a\})$. In this case $a^n \in C_G(a)$ for all $n \in \mathbb{Z}$.

For example, in an abelian group G, $C_G(A) = G$, for all subsets A. One can check by inspection that $C_{Q_8}(i) = \{\pm 1, \pm i\}$. Some other examples are specified in the exercises.

We shall shortly discuss how to minimize the calculation of commutativities between single group elements which appears to be inherent in the computation of centralizers (and other subgroups of a similar nature).

The Center of a Group: Define $Z(G) = \{g \in G \mid gx = xg \text{ for all } x \in G\}$, the set of elements commuting with all the elements of G. This subset of G is called the *center* of G.

Note that $Z(G) = C_G(G)$, so the argument above proves $Z(G) \leq G$ as a special case. As an exercise, the reader may wish to prove $Z(G)$ is a subgroup directly.

Normalizers: Define $gAg^{-1} = \{gag^{-1} \mid a \in A\}$. Define $N_G(A) = \{g \in G \mid gAg^{-1} = A\}$. This subset of G is called the *normalizer* of A in G.

Notice that if $g \in C_G(A)$, then $gag^{-1} = a \in A$ for all $a \in A$ so $C_G(A) \leq N_G(A)$. The proof that $N_G(A)$ is a subgroup of G follows the same steps which demonstrated that $C_G(A) \leq G$ with appropriate modifications.

Examples

(1) If G is abelian then all the elements of G commute, so $Z(G) = G$. Similarly, $C_G(A) = N_G(A) = G$ for *any* subset A of G since $gag^{-1} = gg^{-1}a = a$ for every $g \in G$ and every $a \in A$.

(2) Let $G = D_8$ be the dihedral group of order 8 with the usual generators r and s and let $A = \{1, r, r^2, r^3\}$ be the subgroup of rotations in D_8. We show that $C_{D_8}(A) = A$. Since all powers of r commute with each other, $A \leq C_{D_8}(A)$. Since $sr = rs^{-1} \neq rs$ the element s does not commute with all members of A, i.e. $s \notin C_{D_8}(A)$. Finally, the elements of D_8 that are not in A are all of the form sr^i for some $i \in \{0, 1, 2, 3\}$. If the element sr^i were in $C_{D_8}(A)$ then since $C_{D_8}(A)$ is a *subgroup* which contains r we would also have the element $s = (sr^i)(r^{-i})$ in $C_{D_8}(A)$, a contradiction. This shows $C_{D_8}(A) = A$.

(3) As in the preceding example let $G = D_8$ and let $A = \{1, r, r^2, r^3\}$. We show that $N_{D_8}(A) = D_8$. Since, in general, the centralizer of a subset is contained in its normalizer, $A \leq N_{D_8}(A)$. Next compute that

$$sAs^{-1} = \{s1s^{-1},\ srs^{-1},\ sr^2s^{-1},\ sr^3s^{-1}\} = \{1,\ r^3,\ r^2,\ r\} = A,$$

so that $s \in N_{D_8}(A)$. (Note that the *set* sAs^{-1} equals the *set* A even though the elements in these two sets appear in different orders — this is because s is in the normalizer of A but not in the centralizer of A.) Now both r and s belong to the *subgroup* $N_{D_8}(A)$ and hence $s^i r^j \in N_{D_8}(A)$ for all integers i and j, that is, every element of D_8 is in $N_{D_8}(A)$ (recall that r and s *generate* D_8). Since $D_8 \leq N_{D_8}(A)$ we have $N_{D_8}(A) = D_8$ (the reverse containment being obvious from the definition of a normalizer).

(4) We show that the center of D_8 is the subgroup $\{1, r^2\}$. First observe that the center of any group G is contained in $C_G(A)$ for any subset A of G. Thus by Example 2

above $Z(D_8) \leq C_{D_8}(A) = A$, where $A = \{1, r, r^2, r^3\}$. The calculation in Example 2 shows that r and similarly r^3 are not in $Z(D_8)$, so $Z(D_8) \leq \{1, r^2\}$. To show the reverse inclusion note that r commutes with r^2 and calculate that s also commutes with r^2. Since r and s generate D_8, every element of D_8 commutes with r^2 (and 1), hence $\{1, r^2\} \leq Z(D_8)$ and so equality holds.

(5) Let $G = S_3$ and let A be the subgroup $\{1, (1\,2)\}$. We explain why $C_{S_3}(A) = N_{S_3}(A) = A$. One can compute directly that $C_{S_3}(A) = A$, using the ideas in Example 2 above to minimize the calculations. Alternatively, since an element commutes with its powers, $A \leq C_{S_3}(A)$. By Lagrange's Theorem (Exercise 19 in Section 1.7) the order of the subgroup $C_{S_3}(A)$ of S_3 divides $|S_3| = 6$. Also by Lagrange's Theorem applied to the subgroup A of the group $C_{S_3}(A)$ we have that $2 \mid |C_{S_3}(A)|$. The only possibilities are: $|C_{S_3}(A)| = 2$ or 6. If the latter occurs, $C_{S_3}(A) = S_3$, i.e. $A \leq Z(S_3)$; this is a contradiction because $(1\,2)$ does not commute with $(1\,2\,3)$. Thus $|C_{S_3}(A)| = 2$ and so $A = C_{S_3}(A)$.
Next note that $N_{S_3}(A) = A$ because $\sigma \in N_{S_3}(A)$ if and only if

$$\{\sigma 1 \sigma^{-1}, \sigma(1\,2)\sigma^{-1}\} = \{1, (1\,2)\}.$$

Since $\sigma 1 \sigma^{-1} = 1$, this equality of sets occurs if and only if $\sigma(1\,2)\sigma^{-1} = (1\,2)$ as well, i.e. if and only if $\sigma \in C_{S_3}(A)$.
The center of S_3 is the identity because $Z(S_3) \leq C_{S_3}(A) = A$ and $(1\,2) \notin Z(S_3)$.

Stabilizers and Kernels of Group Actions: The fact that the normalizer of A in G, the centralizer of A in G, and the center of G are all subgroups can be deduced as special cases of results on group actions, indicating that the structure of G is reflected by the sets on which it acts, as follows: if G is a group acting on a set S and s is some fixed element of S, the *stabilizer* of s in G is the set

$$G_s = \{g \in G \mid g \cdot s = s\}$$

(see Exercise 4 in Section 1.7). We show briefly that $G_s \leq G$: first $1 \in G_s$ by axiom (2) of an action. Also, if $y \in G_s$,

$$\begin{aligned} s = 1 \cdot s &= (y^{-1}y) \cdot s \\ &= y^{-1} \cdot (y \cdot s) && \text{(by axiom (1) of an action)} \\ &= y^{-1} \cdot s && \text{(since } y \in G_s \text{)} \end{aligned}$$

so $y^{-1} \in G_s$ as well. Finally, if $x, y \in G_s$, then

$$\begin{aligned} (xy) \cdot s &= x \cdot (y \cdot s) && \text{(by axiom (1) of an action)} \\ &= x \cdot s && \text{(since } y \in G_s \text{)} \\ &= s && \text{(since } x \in G_s \text{) .} \end{aligned}$$

This proves G_s is a subgroup[1] of G. A similar (but easier) argument proves that the *kernel* of an action is a subgroup, where the kernel of the action of G on S is de-

[1] Notice how the steps to prove G_s is a subgroup are the same as those to prove $C_G(A) \leq G$ with axiom (1) of an action taking the place of the associative law.

fined as

$$\{g \in G \mid g \cdot s = s, \text{ for all } s \in S\}$$

(see Exercise 1 in Section 1.7).

Examples

(1) The group $G = D_8$ acts on the set A of four vertices of a square (cf. Example 4 in Section 1.7). The stabilizer of any vertex a is the subgroup $\{1, t\}$ of D_8, where t is the reflection about the line of symmetry passing through vertex a and the center of the square. The kernel of this action is the identity subgroup since only the identity symmetry fixes every vertex.

(2) The group $G = D_8$ also acts on the set A whose elements are the two unordered pairs of opposite vertices (in the labelling of Figure 2 in Section 1.2, $A = \{\ \{1,3\}, \{2,4\}\ \}$). The kernel of the action of D_8 on this set A is the subgroup $\{1, s, r^2, sr^2\}$ and for either element $a \in A$ the stabilizer of a in D_8 equals the kernel of the action.

Finally, we observe that the fact that centralizers, normalizers and kernels are subgroups is a special case of the facts that stabilizers and kernels of actions are subgroups (this will be discussed further in Chapter 4). Let $S = \mathcal{P}(G)$, the collection of all subsets of G, and let G act on S by *conjugation*, that is, for each $g \in G$ and each $B \subseteq G$ let

$$g : B \mapsto gBg^{-1} \quad \text{where} \quad gBg^{-1} = \{gbg^{-1} \mid b \in B\}$$

(see Exercise 16 in Section 1.7). Under this action, it is easy to check that $N_G(A)$ is precisely the stabilizer of A in G (i.e. $N_G(A) = G_s$ where $s = A \in \mathcal{P}(G)$), so $N_G(A)$ is a subgroup of G.

Next let the group $N_G(A)$ act on the set $S = A$ by conjugation, i.e. for all $g \in N_G(A)$ and $a \in A$

$$g : a \mapsto gag^{-1}.$$

Note that this does map A to A by the definition of $N_G(A)$ and so gives an action on A. Here it is easy to check that $C_G(A)$ is precisely the kernel of this action, hence $C_G(A) \leq N_G(A)$; by transitivity of the relation "$\leq$", $C_G(A) \leq G$. Finally, $Z(G)$ is the kernel of G acting on $S = G$ by conjugation, so $Z(G) \leq G$.

EXERCISES

1. Prove that $C_G(A) = \{g \in G \mid g^{-1}ag = a \text{ for all } a \in A\}$.
2. Prove that $C_G(Z(G)) = G$ and deduce that $N_G(Z(G)) = G$.
3. Prove that if A and B are subsets of G with $A \subseteq B$ then $C_G(B)$ is a subgroup of $C_G(A)$.
4. For each of S_3, D_8, and Q_8 compute the centralizers of each element and find the center of each group. Does Lagrange's Theorem (Exercise 19 in Section 1.7) simplify your work?
5. In each of parts (a) to (c) show that for the specified group G and subgroup A of G, $C_G(A) = A$ and $N_G(A) = G$.
   (a) $G = S_3$ and $A = \{1, (1\,2\,3), (1\,3\,2)\}$.
   (b) $G = D_8$ and $A = \{1, s, r^2, sr^2\}$.
   (c) $G = D_{10}$ and $A = \{1, r, r^2, r^3, r^4\}$.

6. Let H be a subgroup of the group G.
   (a) Show that $H \leq N_G(H)$. Give an example to show that this is not necessarily true if H is not a subgroup.
   (b) Show that $H \leq C_G(H)$ if and only if H is abelian.
7. Let $n \in \mathbb{Z}$ with $n \geq 3$. Prove the following:
   (a) $Z(D_{2n}) = 1$ if n is odd
   (b) $Z(D_{2n}) = \{1, r^k\}$ if $n = 2k$.
8. Let G = S_n, fix an $i \in \{1, 2, \ldots, n\}$ and let $G_i = \{\sigma \in G \mid \sigma(i) = i\}$ (the stabilizer of i in G). Use group actions to prove that G_i is a subgroup of G. Find $|G_i|$.
9. For any subgroup H of G and any nonempty subset A of G define $N_H(A) = \{h \in H \mid hAh^{-1} = A\}$. Show that $N_H(A) = N_G(A) \cap H$ and deduce that $N_H(A)$ is a subgroup of H (note that A need not be a subset of H).
10. Let H be a subgroup of order 2 in G. Show that $N_G(H) = C_G(H)$. Deduce that if $N_G(H) = G$ then $H \leq Z(G)$.
11. Prove that $Z(G) \leq N_G(A)$ for any subset A of G.
12. Let R be the set of all polynomials with integer coefficients in the independent variables x_1, x_2, x_3, x_4, i.e. the members of R are finite sums of elements of the form $ax_1^{r_1}x_2^{r_2}x_3^{r_3}x_4^{r_4}$, where a is any integer and $r_1, \ldots, r_4$ are nonnegative integers. For example,

$$12x_1^5x_2^7x_4 - 18x_2^3x_3 + 11x_1^6x_2x_3^3x_4^{23} \qquad (*)$$

is a typical element of R. Each $\sigma \in S_4$ gives a permutation of $\{x_1, \ldots, x_4\}$ by defining $\sigma \cdot x_i = x_{\sigma(i)}$. This may be extended to a map from R to R by defining

$$\sigma \cdot p(x_1, x_2, x_3, x_4) = p(x_{\sigma(1)}, x_{\sigma(2)}, x_{\sigma(3)}, x_{\sigma(4)})$$

for all $p(x_1, x_2, x_3, x_4) \in R$ (i.e. σ simply permutes the indices of the variables). For example, if $\sigma = (1\ 2)(3\ 4)$ and $p(x_1, \ldots, x_4)$ is the polynomial in $(*)$ above, then

$$\begin{aligned}\sigma \cdot p(x_1, x_2, x_3, x_4) &= 12x_2^5x_1^7x_3 - 18x_1^3x_4 + 11x_2^6x_1x_4^3x_3^{23} \\ &= 12x_1^7x_2^5x_3 - 18x_1^3x_4 + 11x_1x_2^6x_3^{23}x_4^3.\end{aligned}$$

   (a) Let $p = p(x_1, \ldots, x_4)$ be the polynomial in $(*)$ above, let $\sigma = (1\ 2\ 3\ 4)$ and let $\tau = (1\ 2\ 3)$. Compute: $\sigma \cdot p$, $\tau \cdot (\sigma \cdot p)$, $(\tau \circ \sigma) \cdot p$, and $(\sigma \circ \tau) \cdot p$.
   (b) Prove that these definitions give a (left) group action of S_4 on R.
   (c) Exhibit all permutations in S_4 that stabilize x_4 and prove that they form a subgroup isomorphic to S_3.
   (d) Exhibit all permutations in S_4 that stabilize the element $x_1 + x_2$ and prove that they form an abelian subgroup of order 4.
   (e) Exhibit all permutations in S_4 that stabilize the element $x_1x_2 + x_3x_4$ and prove that they form a subgroup isomorphic to the dihedral group of order 8.
   (f) Show that the permutations in S_4 that stabilize the element $(x_1 + x_2)(x_3 + x_4)$ are exactly the same as those found in part (e). (The two polynomials appearing in parts (e) and (f) and the subgroup that stabilizes them will play an important role in the study of roots of quartic equations in Section 14.6.)
13. Let n be a positive integer and let R be the set of all polynomials with integer coefficients in the independent variables $x_1, x_2, \ldots, x_n$, i.e. the members of R are finite sums of elements of the form $ax_1^{r_1}x_2^{r_2}\cdots x_n^{r_n}$, where a is any integer and $r_1, \ldots, r_n$ are nonnegative integers.

For each $\sigma \in S_n$ define a map

$$\sigma : R \to R \qquad \text{by} \qquad \sigma \cdot p(x_1, x_2, \ldots, x_n) = p(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}).$$

Prove that this defines a (left) group action of S_n on R.

2.3 CYCLIC GROUPS AND CYCLIC SUBGROUPS

Let G be any group and let x be any element of G. One way of forming a subgroup H of G is by letting H be the set of all integer (positive, negative and zero) powers of x (this guarantees closure under inverses and products at least as far as x is concerned). In this section we study groups which are generated by one element.

Definition. A group H is *cyclic* if there is some element $x \in H$ such that $H = \{x^n \mid n \in \mathbb{Z}\}$ (as usual the operation is multiplication).

In additive notation H is cyclic if $H = \{\, nx \mid n \in \mathbb{Z}\}$. In both cases we shall write $H = \langle x \rangle$ and say H is *generated* by x (and x is a *generator* of H). A cyclic group may have more than one generator. For example, if $H = \langle x \rangle$, then also $H = \langle x^{-1} \rangle$ because $(x^{-1})^n = x^{-n}$ and as n runs over all integers so does $-n$ so that

$$\{x^n \mid n \in \mathbb{Z}\} = \{(x^{-1})^n \mid n \in \mathbb{Z}\}.$$

We shall shortly show how to determine all generators for a given cyclic group H. One should note that the elements of $\langle x \rangle$ are powers of x (or multiples of x, in groups written additively) and not integers. It is not necessarily true that all powers of x are distinct. Also, by the laws for exponents (Exercise 19 in Section 1.1) cyclic groups are abelian.

Examples

(1) Let $G = D_{2n} = \langle r, s \mid r^n = s^2 = 1, rs = sr^{-1} \rangle$, $n \geq 3$ and let H be the subgroup of all rotations of the n-gon. Thus $H = \langle r \rangle$ and the distinct elements of H are $1, r, r^2, \ldots, r^{n-1}$ (these are all the distinct powers of r). In particular, $|H| = n$ and the generator, r, of H has order n. The powers of r "cycle" (forward and backward) with period n, that is,

$$r^n = 1,\ r^{n+1} = r,\ r^{n+2} = r^2, \ldots$$
$$r^{-1} = r^{n-1},\ r^{-2} = r^{n-2}, \ldots \quad etc.$$

In general, to write any power of r, say r^t, in the form r^k, for some k between 0 and $n-1$ use the Division Algorithm to write

$$t = nq + k, \qquad \text{where } 0 \leq k < n,$$

so that

$$r^t = r^{nq+k} = (r^n)^q r^k = 1^q r^k = r^k.$$

For example, in D_8, $r^4 = 1$ so $r^{105} = r^{4(26)+1} = r$ and $r^{-42} = r^{4(-11)+2} = r^2$. Observe that D_{2n} itself is not a cyclic group since it is non-abelian.

(2) Let $H = \mathbb{Z}$ with operation $+$. Thus $H = \langle 1 \rangle$ (here 1 is the integer 1 and the identity of H is 0) and each element in H can be written uniquely in the form $n \cdot 1$, for some $n \in \mathbb{Z}$. In contrast to the preceding example, multiples of the generator are all distinct and we need to take both positive, negative and zero multiples of the generator to obtain all elements of H. In this example $|H|$ and the order of the generator 1 are both ∞. Note also that $H = \langle -1 \rangle$ since each integer x can be written (uniquely) as $(-x)(-1)$.

Before discussing cyclic groups further we prove that the various properties of finite and infinite cyclic groups we observed in the preceding two examples are generic. This proposition also validates the claim (in Chapter 1) that the use of the terminology for "order" of an element and the use of the symbol $|\ \ |$ are consistent with the notion of order of a set.

Proposition 2. If $H = \langle x \rangle$, then $|H| = |x|$ (where if one side of this equality is infinite, so is the other). More specifically

(1) if $|H| = n < \infty$, then $x^n = 1$ and $1, x, x^2, \ldots, x^{n-1}$ are all the distinct elements of H, and

(2) if $|H| = \infty$, then $x^n \neq 1$ for all $n \neq 0$ and $x^a \neq x^b$ for all $a \neq b$ in $\mathbb{Z}$.

Proof: Let $|x| = n$ and first consider the case when $n < \infty$. The elements $1, x, x^2, \ldots, x^{n-1}$ are distinct because if $x^a = x^b$, with, say, $0 \leq a < b < n$, then $x^{b-a} = x^0 = 1$, contrary to n being the smallest positive power of x giving the identity. Thus H has at least n elements and it remains to show that these are all of them. As we did in Example 1, if x^t is any power of x, use the Division Algorithm to write $t = nq + k$, where $0 \leq k < n$, so

$$x^t = x^{nq+k} = (x^n)^q x^k = 1^q x^k = x^k \in \{1, x, x^2, \ldots, x^{n-1}\},$$

as desired.

Next suppose $|x| = \infty$ so no positive power of x is the identity. If $x^a = x^b$, for some a and b with, say, $a < b$, then $x^{b-a} = 1$, a contradiction. Distinct powers of x are distinct elements of H so $|H| = \infty$. This completes the proof of the proposition.

Note that the proof of the proposition gives the method for reducing arbitrary powers of a generator in a finite cyclic group to the "least residue" powers. It is not a coincidence that the calculations of distinct powers of a generator of a cyclic group of order n are carried out via arithmetic in $\mathbb{Z}/n\mathbb{Z}$. Theorem 4 following proves that these two groups are isomorphic.

First we need an easy proposition.

Proposition 3. Let G be an arbitrary group, $x \in G$ and let $m, n \in \mathbb{Z}$. If $x^n = 1$ and $x^m = 1$, then $x^d = 1$, where $d = (m, n)$. In particular, if $x^m = 1$ for some $m \in \mathbb{Z}$, then $|x|$ divides m.

Proof: By the Euclidean Algorithm (see Section 0.2 (6)) there exist integers r and s such that $d = mr + ns$, where d is the g.c.d. of m and n. Thus

$$x^d = x^{mr+ns} = (x^m)^r(x^n)^s = 1^r 1^s = 1.$$

This proves the first assertion.

If $x^m = 1$, let $n = |x|$. If $m = 0$, certainly $n \mid m$, so we may assume $m \neq 0$. Since some nonzero power of x is the identity, $n < \infty$. Let $d = (m, n)$ so by the preceding result $x^d = 1$. Since $0 < d \leq n$ and n is the smallest positive power of x which gives the identity, we must have $d = n$, that is, $n \mid m$, as asserted.

Theorem 4. Any two cyclic groups of the same order are isomorphic. More specifically,

(1) if $n \in \mathbb{Z}^+$ and $\langle x \rangle$ and $\langle y \rangle$ are both cyclic groups of order n, then the map

$$\begin{aligned} \varphi : \langle x \rangle &\to \langle y \rangle \\ x^k &\mapsto y^k \end{aligned}$$

is well defined and is an isomorphism

(2) if $\langle x \rangle$ is an infinite cyclic group, the map

$$\begin{aligned} \varphi : \mathbb{Z} &\to \langle x \rangle \\ k &\mapsto x^k \end{aligned}$$

is well defined and is an isomorphism.

Proof: Suppose $\langle x \rangle$ and $\langle y \rangle$ are both cyclic groups of order n. Let $\varphi : \langle x \rangle \to \langle y \rangle$ be defined by $\varphi(x^k) = y^k$; we must first prove φ is well defined, that is,

$$\text{if } x^r = x^s, \text{ then } \varphi(x^r) = \varphi(x^s).$$

Since $x^{r-s} = 1$, Proposition 3 implies $n \mid r - s$. Write $r = tn + s$ so

$$\begin{aligned} \varphi(x^r) &= \varphi(x^{tn+s}) \\ &= y^{tn+s} \\ &= (y^n)^t y^s \\ &= y^s = \varphi(x^s). \end{aligned}$$

This proves φ is well defined. It is immediate from the laws of exponents that $\varphi(x^a x^b) = \varphi(x^a)\varphi(x^b)$ (check this), that is, φ is a homomorphism. Since the element y^k of $\langle y \rangle$ is the image of x^k under φ, this map is surjective. Since both groups have the same finite order, any surjection from one to the other is a bijection, so φ is an isomorphism (alternatively, φ has an obvious two-sided inverse).

If $\langle x \rangle$ is an infinite cyclic group, let $\varphi : \mathbb{Z} \to \langle x \rangle$ be defined by $\varphi(k) = x^k$. Note that this map is already well defined since there is no ambiguity in the representation of elements in the domain. Since (by Proposition 2) $x^a \neq x^b$, for all distinct $a, b \in \mathbb{Z}$, φ is injective. By definition of a cyclic group, φ is surjective. As above, the laws of exponents ensure φ is a homomorphism, hence φ is an isomorphism, completing the proof.

We chose to use the rotation group $\langle r \rangle$ as our prototypical example of a finite cyclic group of order n (instead of the isomorphic group $\mathbb{Z}/n\mathbb{Z}$) since we shall usually write our cyclic groups multiplicatively:

Notation: For each $n \in \mathbb{Z}^+$, let Z_n be the cyclic group of order n (written multiplicatively).

Up to isomorphism, Z_n is the unique cyclic group of order n and $Z_n \cong \mathbb{Z}/n\mathbb{Z}$. On occasion when we find additive notation advantageous we shall use the latter group as our representative of the isomorphism class of cyclic groups of order n. We shall occasionally say: let $\langle x \rangle$ be the infinite cyclic group (written multiplicatively), however we shall always use $\mathbb{Z}$ (additively) to represent the infinite cyclic group.

As noted earlier, a given cyclic group may have more than one generator. The next two propositions determine precisely which powers of x generate the group $\langle x \rangle$.

Proposition 5. Let G be a group, let $x \in G$ and let $a \in \mathbb{Z} - \{0\}$.

(1) If $|x| = \infty$, then $|x^a| = \infty$.

(2) If $|x| = n < \infty$, then $|x^a| = \dfrac{n}{(n,a)}$.

(3) In particular, if $|x| = n < \infty$ and a is a positive integer dividing n, then $|x^a| = \dfrac{n}{a}$.

Proof: (1) By way of contradiction assume $|x| = \infty$ but $|x^a| = m < \infty$. By definition of order

$$1 = (x^a)^m = x^{am}.$$

Also,

$$x^{-am} = (x^{am})^{-1} = 1^{-1} = 1.$$

Now one of am or $-am$ is positive (since neither a nor m is 0) so some positive power of x is the identity. This contradicts the hypothesis $|x| = \infty$, so the assumption $|x^a| < \infty$ must be false, that is, (1) holds.

(2) Under the notation of (2) let

$$y = x^a, \quad (n,a) = d \quad \text{and write} \quad n = db,\ a = dc,$$

for suitable $b, c \in \mathbb{Z}$ with $b > 0$. Since d is the greatest common divisor of n and a, the integers b and c are relatively prime:

$$(b,c) = 1.$$

To establish (2) we must show $|y| = b$. First note that

$$y^b = x^{ab} = x^{dcb} = (x^{db})^c = (x^n)^c = 1^c = 1$$

so, by Proposition 3 applied to $\langle y \rangle$, we see that $|y|$ divides b. Let $k = |y|$. Then

$$x^{ak} = y^k = 1$$

so by Proposition 3 applied to $\langle x \rangle$, $n \mid ak$, i.e. $db \mid dck$. Thus $b \mid ck$. Since b and c have no factors in common, b must divide k. Since b and k are positive integers which divide each other, $b = k$, which proves (2).

(3) This is a special case of (2) recorded for future reference.

Proposition 6. Let $H = \langle x \rangle$.

(1) Assume $|x| = \infty$. Then $H = \langle x^a \rangle$ if and only if $a = \pm 1$.

(2) Assume $|x| = n < \infty$. Then $H = \langle x^a \rangle$ if and only if $(a, n) = 1$. In particular, the number of generators of H is $\varphi(n)$ (where φ is Euler's φ-function).

Proof: We leave (1) as an exercise. In (2) if $|x| = n < \infty$, Proposition 2 says x^a generates a subgroup of H of order $|x^a|$. This subgroup equals all of H if and only if $|x^a| = |x|$. By Proposition 4,

$$|x^a| = |x| \quad \text{if and only if} \quad \frac{n}{(a,n)} = n, \quad \text{i.e. if and only if } (a,n) = 1.$$

Since $\varphi(n)$ is, by definition, the number of $a \in \{1, 2, \ldots, n\}$ such that $(a, n) = 1$, this is the number of generators of H.

Example

Proposition 6 tells precisely which residue classes mod n generate $\mathbb{Z}/n\mathbb{Z}$: namely, $\bar{a}$ generates $\mathbb{Z}/n\mathbb{Z}$ if and only if $(a, n) = 1$. For instance, $\bar{1}, \bar{5}, \bar{7}$ and $\overline{11}$ are the generators of $\mathbb{Z}/12\mathbb{Z}$ and $\varphi(12) = 4$.

The final theorem in this section gives the complete subgroup structure of a cyclic group.

Theorem 7. Let $H = \langle x \rangle$ be a cyclic group.

(1) Every subgroup of H is cyclic. More precisely, if $K \leq H$, then either $K = \{1\}$ or $K = \langle x^d \rangle$, where d is the smallest positive integer such that $x^d \in K$.

(2) If $|H| = \infty$, then for any distinct nonnegative integers a and b, $\langle x^a \rangle \neq \langle x^b \rangle$. Furthermore, for every integer m, $\langle x^m \rangle = \langle x^{|m|} \rangle$, where $|m|$ denotes the absolute value of m, so that the nontrivial subgroups of H correspond bijectively with the integers $1, 2, 3, \ldots$.

(3) If $|H| = n < \infty$, then for each positive integer a dividing n there is a unique subgroup of H of order a. This subgroup is the cyclic group $\langle x^d \rangle$, where $d = \dfrac{n}{a}$. Furthermore, for every integer m, $\langle x^m \rangle = \langle x^{(n,m)} \rangle$, so that the subgroups of H correspond bijectively with the positive divisors of n.

Proof: (1) Let $K \leq H$. If $K = \{1\}$, the proposition is true for this subgroup, so we assume $K \neq \{1\}$. Thus there exists some $a \neq 0$ such that $x^a \in K$. If $a < 0$ then since K is a group also $x^{-a} = (x^a)^{-1} \in K$. Hence K always contains some positive power of x. Let

$$\mathcal{P} = \{b \mid b \in \mathbb{Z}^+ \text{ and } x^b \in K\}.$$

By the above, $\mathcal{P}$ is a nonempty set of positive integers. By the Well Ordering Principle (Section 0.2) $\mathcal{P}$ has a minimum element—call it d. Since K is a subgroup and $x^d \in K$, $\langle x^d \rangle \leq K$. Since K is a subgroup of H, any element of K is of the form x^a for some integer a. By the Division Algorithm write

$$a = qd + r \qquad 0 \leq r < d.$$

Then $x^r = x^{(a-qd)} = x^a(x^d)^{-q}$ is an element of K since both x^a and x^d are elements of K. By the minimality of d it follows that $r = 0$, i.e. $a = qd$ and so $x^a = (x^d)^q \in \langle x^d \rangle$. This gives the reverse containment $K \leq \langle x^d \rangle$ which proves (1).

We leave the proof of (2) as an exercise (the reasoning is similar to and easier than the proof of (3) which follows).

(3) Assume $|H| = n < \infty$ and $a \mid n$. Let $d = \dfrac{n}{a}$ and apply Proposition 5(3) to obtain that $\langle x^d \rangle$ is a subgroup of order a, showing the existence of a subgroup of order a. To show uniqueness, suppose K is any subgroup of H of order a. By part (1) we have

$$K = \langle x^b \rangle$$

where b is the smallest positive integer such that $x^b \in K$. By Proposition 5

$$\frac{n}{d} = a = |K| = |x^b| = \frac{n}{(n,b)},$$

so $d = (n, b)$. In particular, $d \mid b$. Since b is a multiple of d, $x^b \in \langle x^d \rangle$, hence

$$K = \langle x^b \rangle \leq \langle x^d \rangle.$$

Since $|\langle x^d \rangle| = a = |K|$, we have $K = \langle x^d \rangle$.

The final assertion of (3) follows from the observation that $\langle x^m \rangle$ is a subgroup of $\langle x^{(n,m)} \rangle$ (check this) and, it follows from Proposition 5(2) and Proposition 2 that they have the same order. Since (n, m) is certainly a divisor of n, this shows that every subgroup of H arises from a divisor of n, completing the proof.

Examples

(1) We can use Proposition 6 and Theorem 7 to list all the subgroups of $\mathbb{Z}/n\mathbb{Z}$ for any given n. For example, the subgroups of $\mathbb{Z}/12\mathbb{Z}$ are
(a) $\mathbb{Z}/12\mathbb{Z} = \langle \bar{1} \rangle = \langle \bar{5} \rangle = \langle \bar{7} \rangle = \langle \overline{11} \rangle$ (order 12)
(b) $\langle \bar{2} \rangle = \langle \overline{10} \rangle$ (order 6)
(c) $\langle \bar{3} \rangle = \langle \bar{9} \rangle$ (order 4)
(d) $\langle \bar{4} \rangle = \langle \bar{8} \rangle$ (order 3)
(e) $\langle \bar{6} \rangle$ (order 2)
(f) $\langle \bar{0} \rangle$ (order 1).
The inclusions between them are given by:

$$\langle \bar{a} \rangle \leq \langle \bar{b} \rangle \quad \text{if and only if } (b, 12) \mid (a, 12), \quad 1 \leq a, b \leq 12.$$

(2) We can also combine the results of this section with those of the preceding one. For example, we can obtain subgroups of a group G by forming $C_G(\langle x \rangle)$ and $N_G(\langle x \rangle)$, for each $x \in G$. One can check that an element g in G commutes with x if and only if g commutes with all powers of x, hence

$$C_G(\langle x \rangle) = C_G(x).$$

As noted in Exercise 6, Section 2, $\langle x \rangle \leq N_G(\langle x \rangle)$ but equality need not hold. For instance, if $G = Q_8$ and $x = i$,

$$C_G(\langle i \rangle) = \{\pm 1, \pm i\} = \langle i \rangle \quad \text{and} \quad N_G(\langle i \rangle) = Q_8.$$

Note that we already observed the first of the above two equalities and the second is most easily computed using the result of Exercise 24 following.

EXERCISES

1. Find all subgroups of $Z_{45} = \langle x \rangle$, giving a generator for each. Describe the containments between these subgroups.
2. If x is an element of the finite group G and $|x| = |G|$, prove that $G = \langle x \rangle$. Give an explicit example to show that this result need not be true if G is an infinite group.
3. Find all generators for $\mathbb{Z}/48\mathbb{Z}$.
4. Find all generators for $\mathbb{Z}/202\mathbb{Z}$.
5. Find the number of generators for $\mathbb{Z}/49000\mathbb{Z}$.
6. In $\mathbb{Z}/48\mathbb{Z}$ write out all elements of $\langle \bar{a} \rangle$ for every $\bar{a}$. Find all inclusions between subgroups in $\mathbb{Z}/48\mathbb{Z}$.
7. Let $Z_{48} = \langle x \rangle$ and use the isomorphism $\mathbb{Z}/48\mathbb{Z} \cong Z_{48}$ given by $\bar{1} \mapsto x$ to list all subgroups of Z_{48} as computed in the preceding exercise.
8. Let $Z_{48} = \langle x \rangle$. For which integers a does the map φ_a defined by $\varphi_a : \bar{1} \mapsto x^a$ extend to an *isomorphism* from $\mathbb{Z}/48\mathbb{Z}$ onto Z_{48}.
9. Let $Z_{36} = \langle x \rangle$. For which integers a does the map ψ_a defined by $\psi_a : \bar{1} \mapsto x^a$ extend to a *well defined homomorphism* from $\mathbb{Z}/48\mathbb{Z}$ into Z_{36}. Can ψ_a ever be a surjective homomorphism?
10. What is the order of $\overline{30}$ in $\mathbb{Z}/54\mathbb{Z}$? Write out all of the elements and their orders in $\langle \overline{30} \rangle$.
11. Find all cyclic subgroups of D_8. Find a proper subgroup of D_8 which is not cyclic.
12. Prove that the following groups are *not* cyclic:
    (a) $Z_2 \times Z_2$
    (b) $Z_2 \times \mathbb{Z}$
    (c) $\mathbb{Z} \times \mathbb{Z}$.
13. Prove that the following pairs of groups are *not* isomorphic:
    (a) $\mathbb{Z} \times Z_2$ and $\mathbb{Z}$
    (b) $\mathbb{Q} \times Z_2$ and $\mathbb{Q}$.
14. Let $\sigma = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12)$. For each of the following integers a compute σ^a: $a = 13, 65, 626, 1195, -6, -81, -570$ and -1211.
15. Prove that $\mathbb{Q} \times \mathbb{Q}$ is not cyclic.
16. Assume $|x| = n$ and $|y| = m$. Suppose that x and y *commute*: $xy = yx$. Prove that $|xy|$ divides the least common multiple of m and n. Need this be true if x and y do *not* commute? Give an example of commuting elements x, y such that the order of xy is not equal to the least common multiple of $|x|$ and $|y|$.
17. Find a presentation for Z_n with one generator.
18. Show that if H is any group and h is an element of H with $h^n = 1$, then there is a unique homomorphism from $Z_n = \langle x \rangle$ to H such that $x \mapsto h$.
19. Show that if H is any group and h is an element of H, then there is a unique homomorphism from $\mathbb{Z}$ to H such that $1 \mapsto h$.
20. Let p be a prime and let n be a positive integer. Show that if x is an element of the group G such that $x^{p^n} = 1$ then $|x| = p^m$ for some $m \leq n$.
21. Let p be an odd prime and let n be a positive integer. Use the Binomial Theorem to show

that $(1+p)^{p^{n-1}} \equiv 1 (\text{mod } p^n)$ but $(1+p)^{p^{n-2}} \not\equiv 1 (\text{mod } p^n)$. Deduce that $1+p$ is an element of order p^{n-1} in the multiplicative group $(\mathbb{Z}/p^n\mathbb{Z})^\times$.

22. Let n be an integer ≥ 3. Use the Binomial Theorem to show that $(1+2^2)^{2^{n-2}} \equiv 1 (\text{mod } 2^n)$ but $(1+2^2)^{2^{n-3}} \not\equiv 1 (\text{mod } 2^n)$. Deduce that 5 is an element of order 2^{n-2} in the multiplicative group $(\mathbb{Z}/2^n\mathbb{Z})^\times$.

23. Show that $(\mathbb{Z}/2^n\mathbb{Z})^\times$ is not cyclic for any $n \geq 3$. [Find two distinct subgroups of order 2.]

24. Let G be a finite group and let $x \in G$.

(a) Prove that if $g \in N_G(\langle x \rangle)$ then $gxg^{-1} = x^a$ for some $a \in \mathbb{Z}$.

(b) Prove conversely that if $gxg^{-1} = x^a$ for some $a \in \mathbb{Z}$ then $g \in N_G(\langle x \rangle)$. [Show first that $gx^kg^{-1} = (gxg^{-1})^k = x^{ak}$ for any integer k, so that $g\langle x \rangle g^{-1} \leq \langle x \rangle$. If x has order n, show the elements gx^ig^{-1}, $i = 0, 1, \ldots, n-1$ are distinct, so that $|g\langle x \rangle g^{-1}| = |\langle x \rangle| = n$ and conclude that $g\langle x \rangle g^{-1} = \langle x \rangle$.]

Note that this cuts down some of the work in computing normalizers of cyclic subgroups since one does not have to check $ghg^{-1} \in \langle x \rangle$ for every $h \in \langle x \rangle$.

25. Let G be a finite group of order n. Prove that the map $x \mapsto x^k$ is surjective if k is relatively prime to n. (For such k each element has a kth root in G. We shall see in Section 3.2 that if k is not relatively prime to the order of G then the map $x \mapsto x^k$ is not surjective.)

26. Let $Z_n = \langle x \rangle$ be a cyclic group of order n and for each integer a let

$$\sigma_a : Z_n \to Z_n \qquad \text{by} \qquad \sigma_a(x) = x^a.$$

(a) Prove that σ_a is an automorphism of Z_n if and only if a and n are relatively prime (automorphisms were introduced in Exercise 20, Section 1.6).

(b) Prove that $\sigma_a = \sigma_b$ if and only if $a \equiv b \,(\text{mod } n)$.

(c) Prove that *every* automorphism of Z_n is equal to σ_a for some integer a.

(d) Prove that $\sigma_a \circ \sigma_b = \sigma_{ab}$. Deduce that the map $\bar{a} \mapsto \sigma_a$ is an isomorphism of $(\mathbb{Z}/n\mathbb{Z})^\times$ onto the automorphism group of Z_n (so $\text{Aut}(Z_n)$ is an abelian group of order $\varphi(n)$).

2.4 SUBGROUPS GENERATED BY SUBSETS OF A GROUP

The method of forming cyclic subgroups of a given group is a special case of the general technique where one forms the subgroup generated by an arbitrary subset of a group. In the case of cyclic subgroups one takes a singleton subset $\{x\}$ of the group G and forms all integral powers of x, which amounts to closing the set $\{x\}$ under the group operation and the process of taking inverses. The resulting subgroup is the smallest subgroup of G which contains the set $\{x\}$ (smallest in the sense that if H is any subgroup which contains $\{x\}$, then H contains $\langle x \rangle$). Another way of saying this is that $\langle x \rangle$ is the unique minimal element of the set of subgroups of G containing x (ordered under inclusion). In this section we investigate analogs of this when $\{x\}$ is replaced by an arbitrary subset of G.

Throughout mathematics the following theme recurs: given an object G (such as a group, field, vector space, etc.) and a subset A of G, is there a unique minimal subobject of G (subgroup, subfield, subspace, etc.) which contains A and, if so, how are the elements of this subobject computed? Students may already have encountered this question in the study of vector spaces. When G is a vector space (with, say, real number scalars) and $A = \{v_1, v_2, \ldots, v_n\}$, then there is a unique smallest subspace of

G which contains A, namely the (linear) span of $v_1, v_2, \ldots, v_n$ and each vector in this span can be written as $k_1v_1 + k_2v_2 + \cdots + k_nv_n$, for some $k_1, \ldots, k_n \in \mathbb{R}$. When A is a single nonzero vector, v, the span of $\{v\}$ is simply the 1-dimensional subspace or line containing v and every element of this subspace is of the form kv for some $k \in \mathbb{R}$. This is the analog in the theory of vector spaces of cyclic subgroups of a group. Note that the 1-dimensional subspaces contain kv, where $k \in \mathbb{R}$, not just kv, where $k \in \mathbb{Z}$; the reason being that a subspace must be closed under *all* the vector space operations (*e.g.* scalar multiplication) not just the group operation of vector addition.

Let G be any group and let A be any subset of G. We now make precise the notion of the subgroup of G generated by A. We prove that because the intersection of any set of subgroups of G is also a subgroup of G, the subgroup generated by A is the unique smallest subgroup of G containing A; it is "smallest" in the sense of being the minimal element of the set of all subgroups containing A. We show that the elements of this subgroup are obtained by closing the given subset under the group operation (and taking inverses). In succeeding parts of the text when we develop the theory of other algebraic objects we shall refer to this section as the paradigm in proving that a given subset is contained in a unique smallest subobject and that the elements of this subobject are obtained by closing the subset under the operations which define the object. Since in the latter chapters the details will be omitted, students should acquire a solid understanding of the process at this point.

In order to proceed we need only the following:

Proposition 8. If $\mathcal{A}$ is any nonempty collection of subgroups of G, then the intersection of all members of $\mathcal{A}$ is also a subgroup of G.

Proof: This is an easy application of the subgroup criterion (see also Exercise 10, Section 1). Let

$$K = \bigcap_{H \in \mathcal{A}} H.$$

Since each $H \in \mathcal{A}$ is a subgroup, $1 \in H$, so $1 \in K$, that is, $K \neq \emptyset$. If $a, b \in K$, then $a, b \in H$, for all $H \in \mathcal{A}$. Since each H is a group, $ab^{-1} \in H$, for all H, hence $ab^{-1} \in K$. Proposition 1 gives that $K \leq G$.

Definition. If A is any subset of the group G define

$$\langle A \rangle = \bigcap_{\substack{A \subseteq H \\ H \leq G}} H.$$

This is called the *subgroup of G generated by A*.

Thus $\langle A \rangle$ is the intersection of all subgroups of G containing A. It is a subgroup of G by Proposition 8 applied to the set $\mathcal{A} = \{H \leq G \mid A \subseteq H\}$ ($\mathcal{A}$ is nonempty since $G \in \mathcal{A}$). Since A lies in each $H \in \mathcal{A}$, A is a subset of their intersection, $\langle A \rangle$. Note that $\langle A \rangle$ is the unique minimal element of $\mathcal{A}$ as follows: $\langle A \rangle$ is a subgroup of G containing A, so $\langle A \rangle \in \mathcal{A}$; and any element of $\mathcal{A}$ contains the intersection of all elements in $\mathcal{A}$, i.e. contains $\langle A \rangle$.

When A is the finite set $\{a_1, a_2, \ldots, a_n\}$ we write $\langle a_1, a_2, \ldots, a_n\rangle$ for the group generated by $a_1, a_2, \ldots, a_n$ instead of $\langle\{a_1, a_2, \ldots, a_n\}\rangle$. If A and B are two subsets of G we shall write $\langle A, B\rangle$ in place of $\langle A \cup B\rangle$.

This "top down" approach to defining $\langle A\rangle$ proves existence and uniqueness of the smallest subgroup of G containing A but is not too enlightening as to how to construct the elements in it. As the word "generates" suggests we now define the set which is the closure of A under the group operation (and the process of taking inverses) and prove this set equals $\langle A\rangle$. Let

$$\overline{A} = \{a_1^{\epsilon_1} a_2^{\epsilon_2} \ldots a_n^{\epsilon_n} \mid n \in \mathbb{Z},\ n \geq 0 \text{ and } a_i \in A, \epsilon_i = \pm 1 \text{ for each } i\}$$

where $\overline{A} = \{1\}$ if $A = \emptyset$, so that $\overline{A}$ is the set of all finite products (called *words*) of elements of A and inverses of elements of A. Note that the a_i's need not be distinct, so a^2 is written aa in the notation defining $\overline{A}$. Note also that A is not assumed to be a finite (or even countable) set.

Proposition 9. $\overline{A} = \langle A\rangle$.

Proof: We first prove $\overline{A}$ is a subgroup. Note that $\overline{A} \neq \emptyset$ (even if $A = \emptyset$). If $a, b \in \overline{A}$ with $a = a_1^{\epsilon_1} a_2^{\epsilon_2} \ldots a_n^{\epsilon_n}$ and $b = b_1^{\delta_1} b_2^{\delta_2} \ldots b_m^{\delta_m}$, then

$$ab^{-1} = a_1^{\epsilon_1} a_2^{\epsilon_2} \ldots a_n^{\epsilon_n} \cdot b_m^{-\delta_m} b_{m-1}^{-\delta_{m-1}} \ldots b_1^{-\delta_1}$$

(where we used Exercise 15 of Section 1.1 to compute b^{-1}). Thus ab^{-1} is a product of elements of A raised to powers ± 1, hence $ab^{-1} \in \overline{A}$. Proposition 1 implies $\overline{A}$ is a subgroup of G.

Since each $a \in A$ may be written a^1, it follows that $A \subseteq \overline{A}$, hence $\langle A\rangle \subseteq \overline{A}$. But $\langle A\rangle$ is a group containing A and, since it is closed under the group operation and the process of taking inverses, $\langle A\rangle$ contains each element of the form $a_1^{\epsilon_1} a_2^{\epsilon_2} \ldots a_n^{\epsilon_n}$, that is, $\overline{A} \subseteq \langle A\rangle$. This completes the proof of the proposition.

We now use $\langle A\rangle$ in place of $\overline{A}$ and may take the definition of $\overline{A}$ as an equivalent definition of $\langle A\rangle$. As noted above, in this equivalent definition of $\langle A\rangle$, products of the form $a \cdot a$, $a \cdot a \cdot a$, $a \cdot a^{-1}$, etc. could have been simplified to a^2, a^3, 1, etc. respectively, so another way of writing $\langle A\rangle$ is

$$\langle A\rangle = \{a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_n^{\alpha_n} \mid \text{for each } i, \quad a_i \in A, \alpha_i \in \mathbb{Z}, a_i \neq a_{i+1} \text{ and } n \in \mathbb{Z}^+\}.$$

In fact, when $A = \{x\}$ this was our definition of $\langle A\rangle$.

If G is *abelian*, we could commute the a_i's and so collect all powers of a given generator together. For instance, if A were the finite subset $\{a_1, a_2, \ldots, a_k\}$ of the abelian group G, one easily checks that

$$\langle A\rangle = \{a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_k^{\alpha_k} \mid \alpha_i \in \mathbb{Z} \text{ for each } i\}.$$

If in this situation we further assume that each a_i has finite order d_i, for all i, then since there are exactly d_i distinct powers of a_i, the total number of distinct products of the form $a_1^{\alpha_1} a_2^{\alpha_2} \ldots a_k^{\alpha_k}$ is at most $d_1 d_2 \ldots d_k$, that is,

$$|\langle A\rangle| \leq d_1 d_2 \ldots d_k.$$

It may happen that $a^\alpha b^\beta = a^\gamma b^\delta$ even though $a^\alpha \neq a^\gamma$ and $b^\beta \neq b^\delta$. We shall explore exactly when this happens when we study direct products in Chapter 5.

When G is *non-abelian* the situation is much more complicated. For example, let $G = D_8$ and let r and s be the usual generators of D_8 (note that the notation $D_8 = \langle r, s\rangle$ is consistent with the notation introduced in Section 1.2). Let $a = s$, let $b = rs$ and let $A = \{a, b\}$. Since both s and r $(= rs \cdot s)$ belong to $\langle a, b\rangle$, $G = \langle a, b\rangle$, i.e. G is also generated by a and b. Both a and b have order 2, however D_8 has order 8. This means that it is *not* possible to write every element of D_8 in the form $a^\alpha b^\beta$, $\alpha, \beta \in \mathbb{Z}$. More specifically, the product aba cannot be simplified to a product of the form $a^\alpha b^\beta$. In fact, if $G = D_{2n}$ for any $n > 2$, and r, s, a, b are defined in the same way as above, it is still true that

$$|a| = |b| = 2, \quad D_{2n} = \langle a, b\rangle \quad \text{and} \quad |D_{2n}| = 2n.$$

This means that for large n, long products of the form $abab\ldots ab$ cannot be further simplified. In particular, this illustrates that, unlike the abelian (or, better yet, cyclic) group case, the order of a (finite) group cannot even be bounded once we know the orders of the elements in some generating set.

Another example of this phenomenon is S_n:

$$S_n = \langle (1\,2), (1\,2\,3\ldots n)\rangle.$$

Thus S_n is generated by an element of order 2 together with one of order n, yet $|S_n| = n!$ (we shall prove these statements later after developing some more techniques).

One final example emphasizes the fact that if G is non-abelian, subgroups of G generated by more than one element of G may be quite complicated. Let

$$G = GL_2(\mathbb{R}), \quad a = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 & 2 \\ 1/2 & 0 \end{pmatrix}$$

so $a^2 = b^2 = 1$ but $ab = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$. It is easy to see that ab has infinite order, so $\langle a, b\rangle$ is an *infinite* subgroup of $GL_2(\mathbb{R})$ which is generated by two elements of order 2.

These examples illustrate that when $|A| \geq 2$ it is difficult, in general, to compute even the order of the subgroup generated by A, let alone any other structural properties. It is therefore impractical to gather much information about subgroups of a non-abelian group created by taking random subsets A and trying to write out the elements of (or other information about) $\langle A\rangle$. For certain "well chosen" subsets A, even of a non-abelian group G, we shall be able to make both theoretical and computational use of the subgroup generated by A. One example of this might be when we want to find a subgroup of G which contains $\langle x\rangle$ properly; we might search for some element y which commutes with x (i.e. $y \in C_G(x)$) and form $\langle x, y\rangle$. It is easy to check that the latter group is abelian, so its order is bounded by $|x||y|$. Alternatively, we might instead take y in $N_G(\langle x\rangle)$ — in this case the same order bound holds and the structure of $\langle x, y\rangle$ is again not too complicated (as we shall see in the next chapter).

The complications which arise for non-abelian groups are generally not quite as serious when we study other basic algebraic systems because of the additional algebraic structure imposed.

EXERCISES

1. Prove that if H is a subgroup of G then $\langle H \rangle = H$.
2. Prove that if A is a subset of B then $\langle A \rangle \leq \langle B \rangle$. Give an example where $A \subseteq B$ with $A \neq B$ but $\langle A \rangle = \langle B \rangle$.
3. Prove that if H is an abelian subgroup of a group G then $\langle H, Z(G) \rangle$ is abelian. Give an explicit example of an abelian subgroup H of a group G such that $\langle H, C_G(H) \rangle$ is not abelian.
4. Prove that if H is a subgroup of G then H is generated by the set $H - \{1\}$.
5. Prove that the subgroup generated by any two distinct elements of order 2 in S_3 is all of S_3.
6. Prove that the subgroup of S_4 generated by $(1\ 2)$ and $(1\ 2)(3\ 4)$ is a noncyclic group of order 4.
7. Prove that the subgroup of S_4 generated by $(1\ 2)$ and $(1\ 3)(2\ 4)$ is isomorphic to the dihedral group of order 8.
8. Prove that $S_4 = \langle (1\ 2\ 3\ 4), (1\ 2\ 4\ 3) \rangle$.
9. Prove that $SL_2(\mathbb{F}_3)$ is the subgroup of $GL_2(\mathbb{F}_3)$ generated by $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$. [Recall from Exercise 9 of Section 1 that $SL_2(\mathbb{F}_3)$ is the subgroup of matrices of determinant 1. You may assume this subgroup has order 24 — this will be an exercise in Section 3.2.]
10. Prove that the subgroup of $SL_2(\mathbb{F}_3)$ generated by $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ is isomorphic to the quaternion group of order 8. [Use a presentation for Q_8.]
11. Show that $SL_2(\mathbb{F}_3)$ and S_4 are two nonisomorphic groups of order 24.
12. Prove that the subgroup of upper triangular matrices in $GL_3(\mathbb{F}_2)$ is isomorphic to the dihedral group of order 8 (cf. Exercise 16, Section 1). [First find the order of this subgroup.]
13. Prove that the multiplicative group $\mathbb{Q}^\times$ of nonzero rational numbers is generated by $\{\frac{1}{p} \mid p \text{ is a prime}\}$.
14. A group H is called *finitely generated* if there is a finite set A such that $H = \langle A \rangle$.
    (a) Prove that every finite group is finitely generated.
    (b) Prove that $\mathbb{Z}$ is finitely generated.
    (c) Prove that every finitely generated subgroup of the additive group $\mathbb{Q}$ is cyclic. [If H is a finitely generated subgroup of $\mathbb{Q}$, show that $H \leq \langle \frac{1}{k} \rangle$, where k is the product of all the denominators which appear in a set of generators for H.]
    (d) Prove that $\mathbb{Q}$ is not finitely generated.
15. Exhibit a proper subgroup of $\mathbb{Q}$ which is not cyclic.
16. A subgroup M of a group G is called a *maximal subgroup* if $M \neq G$ and the only subgroups of G which contain M are M and G.
    (a) Prove that if H is a proper subgroup of the finite group G then there is a maximal subgroup of G containing H.
    (b) Show that the subgroup of all rotations in a dihedral group is a maximal subgroup.
    (c) Show that if $G = \langle x \rangle$ is a cyclic group of order $n \geq 1$ then a subgroup H is maximal if and only if $H = \langle x^p \rangle$ for some prime p dividing n.
17. This is an exercise involving Zorn's Lemma (see the Appendix) to prove that every nontrivial finitely generated group possesses maximal subgroups. Let G be a finitely generated group, say $G = \langle g_1, g_2, \ldots, g_n \rangle$, and let $\mathcal{S}$ be the set of all proper subgroups of G. Then $\mathcal{S}$ is partially ordered by inclusion. Let $\mathcal{C}$ be a chain in $\mathcal{S}$.

(a) Prove that the union, H, of all the subgroups in $\mathcal{C}$ is a subgroup of G.
(b) Prove that H is a *proper* subgroup. [If not, each g_i must lie in H and so must lie in some element of the chain $\mathcal{C}$. Use the definition of a chain to arrive at a contradiction.]
(c) Use Zorn's Lemma to show that $\mathcal{S}$ has a maximal element (which is, by definition, a maximal subgroup).

18. Let p be a prime and let $Z = \{z \in \mathbb{C} \mid z^{p^n} = 1 \text{ for some } n \in \mathbb{Z}^+\}$ (so Z is the multiplicative group of all p-power roots of unity in $\mathbb{C}$). For each $k \in \mathbb{Z}^+$ let $H_k = \{z \in Z \mid z^{p^k} = 1\}$ (the group of p^kth roots of unity). Prove the following:
(a) $H_k \leq H_m$ if and only if $k \leq m$
(b) H_k is cyclic for all k (you may assume that for any $n \in \mathbb{Z}^+$, $\{e^{2\pi i t/n} \mid t = 0, 1, \ldots, n-1\}$ is the set of all n^{th} roots of 1 in $\mathbb{C}$)
(c) every proper subgroup of Z equals H_k for some $k \in \mathbb{Z}^+$ (in particular, every proper subgroup of Z is finite and cyclic)
(d) Z is not finitely generated.

19. A nontrivial abelian group A (written multiplicatively) is called *divisible* if for each element $a \in A$ and each nonzero integer k there is an element $x \in A$ such that $x^k = a$, i.e. each element has a kth root in A (in additive notation, each element is the kth multiple of some element of A).
(a) Prove that the additive group of rational numbers, $\mathbb{Q}$, is divisible.
(b) Prove that no finite abelian group is divisible.

20. Prove that if A and B are nontrivial abelian groups, then $A \times B$ is divisible if and only if both A and B are divisible groups.

2.5 THE LATTICE OF SUBGROUPS OF A GROUP

In this section we describe a graph associated to a group which depicts the relationships between its subgroups. This graph, called the lattice[2] of subgroups of the group, is a good way of "visualizing" a group — it certainly illuminates the structure of a group better than the group table. We shall be using lattice diagrams, or parts of them, to describe both specific groups and certain properties of general groups throughout the chapters on group theory. Moreover, the lattice of subgroups of a group will play an important role in Galois Theory.

The lattice of subgroups of a given finite group G is constructed as follows: plot all subgroups of G starting at the bottom with 1, ending at the top with G and, roughly speaking, with subgroups of larger order positioned higher on the page than those of smaller order. Draw paths upwards between subgroups using the rule that there will be a line upward from A to B if $A \leq B$ and there are no subgroups properly between A and B. Thus if $A \leq B$ there is a path (possibly many paths) upward from A to B passing through a chain of intermediate subgroups (and a path downward from B to A if $B \geq A$). The initial positioning of the subgroups on the page, which is, a priori, somewhat arbitrary, can often (with practice) be chosen to produce a simple picture. Notice that for any pair of subgroups H and K of G the unique smallest subgroup which contains both of them, namely $\langle H, K \rangle$ (called the *join* of H and K),

[2] The term "lattice" has a precise mathematical meaning in terms of partially ordered sets.

may be read off from the lattice as follows: trace paths upwards from H and K until a common subgroup A which contains H and K is reached (note that G itself always contains all subgroups so at least one such A exists). To ensure that $A = \langle H, K \rangle$ make sure there is no $A_1 \leq A$ (indicated by a downward path from A to A_1) with both H and K contained in A_1 (otherwise replace A with A_1 and repeat the process to see if $A_1 = \langle H, K \rangle$). By a symmetric process one can read off the largest subgroup of G which is contained in both H and K, namely their intersection (which is a subgroup by Proposition 8).

There are some limitations to this process, in particular it cannot be carried out *per se* for infinite groups. Even for finite groups of relatively small order, lattices can be quite complicated (see the book: *Groups of Order* 2^n, $n \leq 6$ by M. Hall and J. Senior, Macmillan, 1964, for some hair-raising examples). At the end of this section we shall describe how parts of a lattice may be drawn and used even for infinite groups.

Note that isomorphic groups have the same lattices (i.e. the same directed graphs). Nonisomorphic groups may also have identical lattices (this happens for two groups of order 16—see the following exercises). Since the lattice of subgroups is only part of the data we shall carry in our descriptors of a group, this will not be a serious drawback (indeed, it might even be useful in seeing when two nonisomorphic groups have some common properties).

Examples

Except for the cyclic groups (Example 1) we have not proved that the following lattices are correct (e.g. contain all subgroups of the given group or have the right joins and intersections). For the moment we shall take these facts as given and, as we build up more theory in the course of the text, we shall assign as exercises the proofs that these are indeed correct.

(1) For $G = Z_n \cong \mathbb{Z}/n\mathbb{Z}$, by Theorem 7 the lattice of subgroups of G is the lattice of divisors of n (that is, the divisors of n are written on a page with n at the bottom, 1 at the top and paths upwards from a to b if $b \mid a$). Some specific examples for various values of n follow.

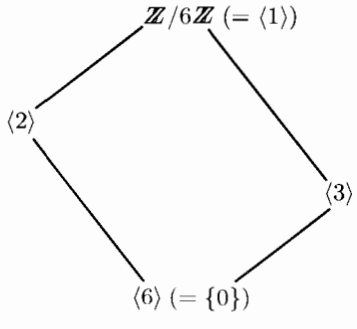
$$\begin{array}{ll} \mathbb{Z}/2\mathbb{Z} & = \langle 1 \rangle \\ \quad | & \\ \langle 2 \rangle & = \{0\} \end{array}$$

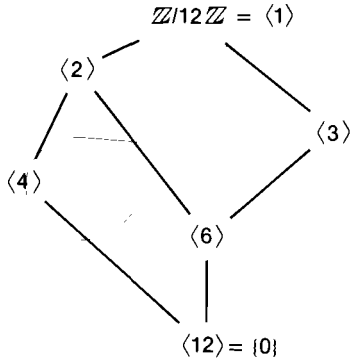
$$\begin{array}{ll} \mathbb{Z}/4\mathbb{Z} & = \langle 1 \rangle \text{ (note: } \langle 1 \rangle = \langle 3 \rangle) \\ \quad | & \\ \langle 2 \rangle & \\ \quad | & \\ \langle 4 \rangle & = \{0\} \end{array}$$

$$
\begin{array}{ll}
\mathbb{Z}/8\mathbb{Z} & = \langle 1 \rangle \text{ (note: } \langle 1 \rangle = \langle 3 \rangle = \langle 5 \rangle = \langle 7 \rangle) \\
| & \\
\langle 2 \rangle & \\
| & \\
\langle 4 \rangle & \\
| & \\
\langle 8 \rangle & = \{0\}
\end{array}
$$

In general, if p is a prime, the lattice of $\mathbb{Z}/p^n\mathbb{Z}$ is:

$$
\begin{array}{ll}
\mathbb{Z}/p^n\mathbb{Z} & = \langle 1 \rangle \\
| & \\
\langle p \rangle & \\
| & \\
\langle p^2 \rangle & \\
| & \\
\langle p^3 \rangle & \\
| & \\
\vdots & \\
| & \\
\langle p^{n-1} \rangle & \\
| & \\
\langle p^n \rangle & = \{0\}
\end{array}
$$

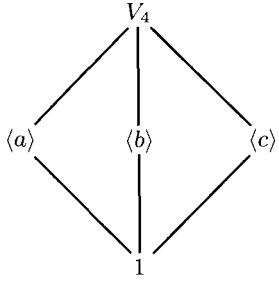




(2) The *Klein 4-group (Viergruppe)*, V_4, is the group of order 4 with multiplication table:

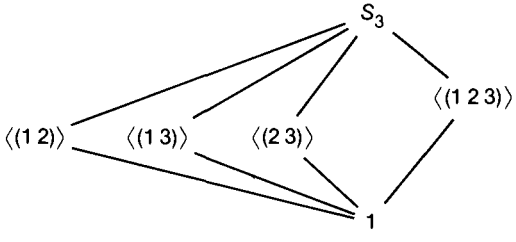
$\cdot$	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

and lattice

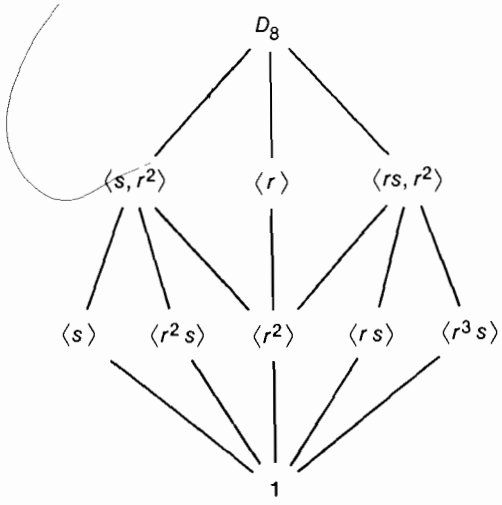


Note that V_4 is abelian and is not isomorphic to Z_4 (why?). We shall see that D_8 has an isomorphic copy of V_4 as a subgroup, so it will not be necessary to check that the associative law holds for the binary operation defined above.

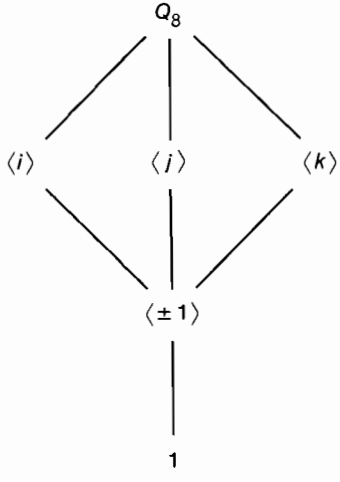
(3) The lattice of S_3 is:



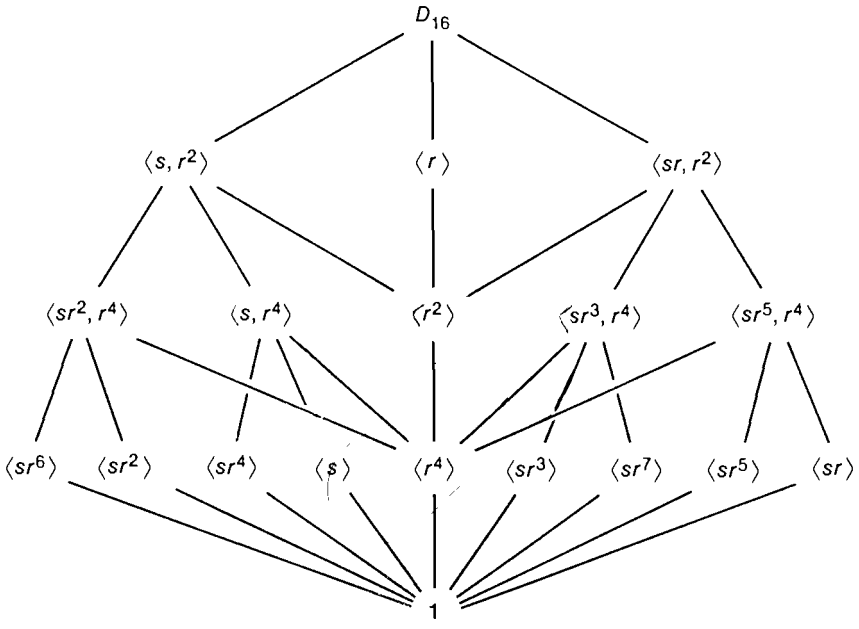
(4) Using our usual notation for $D_8 = \langle r, s \rangle$, the lattice of D_8 is:



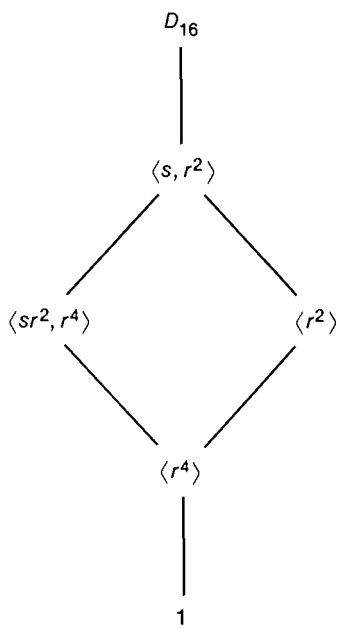
(5) The lattice of subgroups of Q_8 is:



(6) The lattice of D_{16} is not a planar graph (cannot be drawn on a plane without lines crossing). One way of drawing it is:



In many instances in both theoretical proofs and specific examples we shall be only interested in information concerning two (or some small number of) subgroups of a given group and their interrelationships. To depict these graphically we shall draw a *sublattice* of the entire group lattice which contains the relevant joins and intersections. An unbroken line in such a sublattice will not, in general, mean that there is no subgroup in between the endpoints of the line. These partial lattices for groups will also be used when we are dealing with infinite groups. For example, if we wished only to discuss the relationship between the subgroups: $\langle sr^2, r^4 \rangle$ and $\langle r^2 \rangle$ of D_{16} we would draw the sublattice:



Note that $\langle s, r^2 \rangle$ and $\langle r^4 \rangle$ are precisely the join and intersection, respectively, of these two subgroups in D_{16}.

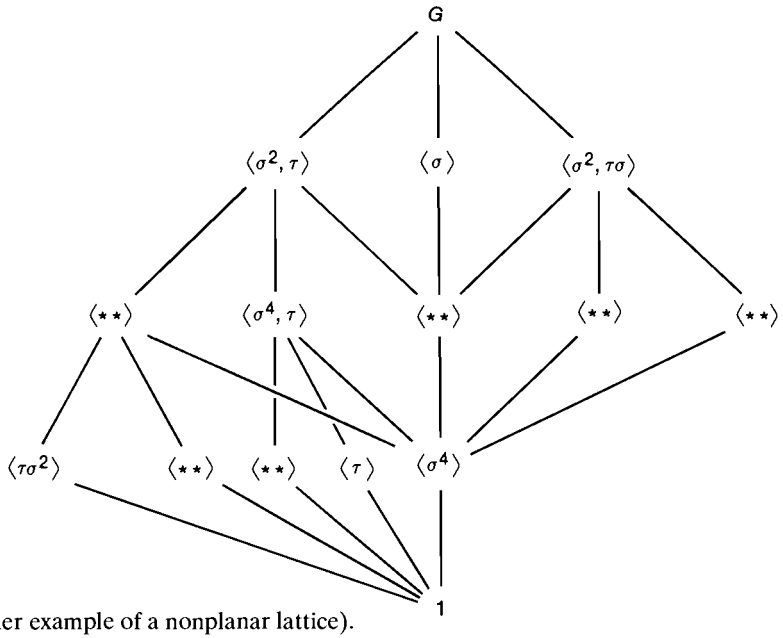
Finally, given the lattice of subgroups of a group, it is relatively easy to compute normalizers and centralizers. For example, in D_8 we can see that $C_{D_8}(s) = \langle s, r^2 \rangle$ because we first calculate that $r^2 \in C_{D_8}(s)$ (see Section 2). This proves $\langle s, r^2 \rangle \leq C_{D_8}(s)$ (note that an element always belongs to its own centralizer). The only subgroups which contain $\langle s, r^2 \rangle$ are that subgroup itself and all of D_8. We cannot have $C_{D_8}(s) = D_8$ because r does not commute with s (i.e. $r \notin C_{D_8}(s)$). This leaves only the claimed possibility for $C_{D_8}(s)$.

EXERCISES

1. Let H and K be subgroups of G. Exhibit all possible sublattices which show only G, 1, H, K and their joins and intersections. What distinguishes the different drawings?
2. In each of (a) to (d) list all subgroups of D_{16} that satisfy the given condition.
   (a) Subgroups that are contained in $\langle sr^2, r^4 \rangle$.
   (b) Subgroups that are contained in $\langle sr^7, r^4 \rangle$.
   (c) Subgroups that contain $\langle r^4 \rangle$.
   (d) Subgroups that contain $\langle s \rangle$.
3. Show that the subgroup $\langle s, r^2 \rangle$ of D_8 is isomorphic to V_4.
4. Use the given lattice to find all pairs of elements that generate D_8 (there are 12 pairs).
5. Use the given lattice to find all elements $x \in D_{16}$ such that $D_{16} = \langle x, s \rangle$ (there are 16 such elements x).
6. Use the given lattices to help find the centralizers of every element in the following groups:
   (a) D_8
) Q_8
   (c) S_3
   (d) D_{16}.
7. Find the center of D_{16}.
8. In each of the following groups find the normalizer of each subgroup:
   (a) S_3
   (b) Q_8.
9. Draw the lattices of subgroups of the following groups:
   (a) $\mathbb{Z}/16\mathbb{Z}$
   (b) $\mathbb{Z}/24\mathbb{Z}$
   (c) $\mathbb{Z}/48\mathbb{Z}$. [See Exercise 6 in Section 3.]
10. Classify groups of order 4 by proving: if $|G| = 4$ then $G \cong Z_4$ or $G \cong V_4$. [See Exercise 36, Section 1.1.]
11. Consider the group of order 16 with the following presentation:

$$QD_{16} = \langle \sigma, \tau \mid \sigma^8 = \tau^2 = 1,\ \sigma\tau = \tau\sigma^3 \rangle$$

(called the *quasidihedral* or *semidihedral* group of order 16). This group has three subgroups of order 8: $\langle \tau, \sigma^2 \rangle \cong D_8$, $\langle \sigma \rangle \cong Z_8$ and $\langle \sigma^2, \sigma\tau \rangle \cong Q_8$ and every proper subgroup is contained in one of these three subgroups. Fill in the missing subgroups in the lattice of all subgroups below (each subgroup should be exhibited with at most two generators).



(this is another example of a nonplanar lattice).

The next three examples lead to two nonisomorphic groups that have the same lattice o subgroups.

12. The group $A = Z_2 \times Z_4 = \langle a, b \mid a^2 = b^4 = 1,\ ab = ba\rangle$ has order 8 and has thre subgroups of order 4: $\langle a, b^2\rangle \cong V_4$, $\langle b\rangle \cong Z_4$ and $\langle ab\rangle \cong Z_4$ and every proper subgroup i contained in one of these three. Draw the lattice of all subgroups of A, giving each subgrou in terms of at most two generators.

13. The group $G = Z_2 \times Z_8 = \langle x, y \mid x^2 = y^8 = 1,\ xy = yx\rangle$ has order 16 and has thre subgroups of order 8: $\langle x, y^2\rangle \cong Z_2 \times Z_4$, $\langle y\rangle \cong Z_8$ and $\langle xy\rangle \cong Z_8$ and every prope subgroup is contained in one of these three. Draw the lattice of all subgroups of G, givi each subgroup in terms of at most two generators (cf. Exercise 12).

14. Let M be the group of order 16 with the following presentation:

$$\langle u, v \mid u^2 = v^8 = 1,\ vu = uv^5\rangle$$

(sometimes called the *modular* group of order 16). It has three subgroups of order $\langle u, v^2\rangle$, $\langle v\rangle$ and $\langle uv\rangle$ and every proper subgroup is contained in one of these three. Pr that $\langle u, v^2\rangle \cong Z_2 \times Z_4$, $\langle v\rangle \cong Z_8$ and $\langle uv\rangle \cong Z_8$. Show that the lattice of subgroups M is the same as the lattice of subgroups of $Z_2 \times Z_8$ (cf. Exercise 13) but that these groups are not isomorphic.

15. Describe the isomorphism type of each of the three subgroups of D_{16} of order 8.

16. Use the lattice of subgroups of the quasidihedral group of order 16 to show that e element of order 2 is contained in the proper subgroup $\langle\tau, \sigma^2\rangle$ (cf. Exercise 11).

17. Use the lattice of subgroups of the modular group M of order 16 to show that $\{x \in$ $x^2 = 1\}$ is a subgroup of M isomorphic to the Klein 4-group (cf. Exercise 14).

18. Use the lattice to help find the centralizer of every element of QD_{16} (cf. Exercise 11).

19. Use the lattice to help find $N_{D_{16}}(\langle s, r^4\rangle)$.

20. Use the lattice of subgroups of QD_{16} (cf. Exercise 11) to help to find
(a) $N_{QD_{16}}(\langle\tau\sigma\rangle)$, and
(b) $N_{QD_{16}}(\langle\tau, \sigma^4\rangle)$.

3

Quotient Groups and Homomorphisms

3.1 DEFINITIONS AND EXAMPLES

In this chapter we introduce the notion of a *quotient* group of a group G, which is another way of obtaining a "smaller" group from the group G and, as we did with subgroups, we shall use quotient groups to study the structure of G. The structure of the group G is reflected in the structure of the quotient groups and the subgroups of G. For example, we shall see that the lattice of subgroups for a *quotient* of G is reflected at the "top" (in a precise sense) of the lattice for G whereas the lattice for a *subgroup* of G occurs naturally at the "bottom". One can therefore obtain information about the group G by combining this information and we shall indicate how some classification theorems arise in this way.

The study of the quotient groups of G is essentially equivalent to the study of the homomorphisms of G, i.e. the maps of the group G to another group which respect the group structures. If φ is a homomorphism from G to a group H recall that the *fibers* of φ are the sets of elements of G projecting to single elements of H, which we can represent pictorially in Fig. 1, where the vertical line in the box above a point a represents the fiber of φ over a.

The group operation in H provides a way to multiply two elements in the image of φ (i.e. two elements on the horizontal line in Fig. 1). This suggests a natural multiplication of the *fibers* lying above these two points making *the set of fibers into a group*: if X_a is the fiber above a and X_b is the fiber above b then the product of X_a with X_b is defined to be the fiber X_{ab} above the product ab, i.e. $X_aX_b = X_{ab}$. This multiplication is associative since multiplication is associative in H, the identity is the fiber over the identity of H, and the inverse of the fiber over a is the fiber over a^{-1}, as is easily checked from the definition.[1] Roughly speaking, the group G is partitioned into pieces (the fibers) and these pieces have themselves the structure of a group, called a *quotient* group of G (a formal definition follows the example below).

Since the multiplication of fibers is defined from the multiplication in H, by construction the quotient group with this multiplication is naturally isomorphic to the image of G under the homomorphism φ (the fiber X_a is identified with its image element a in H).

[1] For example, the associativity is proved as follows: $(X_aX_b)X_c = (X_{ab})X_c = X_{(ab)c}$ and $X_a(X_bX_c) = X_a(X_{bc}) = X_{a(bc)}$. Since $(ab)c = a(bc)$ in H, $(X_aX_b)X_c = X_a(X_bX_c)$.

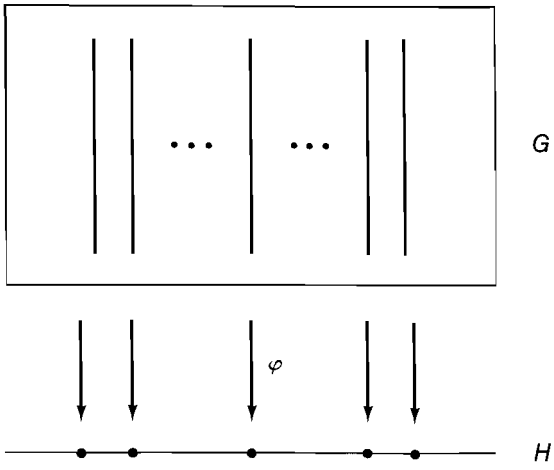


Fig. 1

Example

Let $G = \mathbb{Z}$, let $H = Z_n = \langle x \rangle$ be the cyclic group of order n and define $\varphi : \mathbb{Z} \to Z_n$ by $\varphi(a) = x^a$. Since

$$\varphi(a+b) = x^{a+b} = x^a x^b = \varphi(a)\varphi(b)$$

it follows that φ is a homomorphism (note that the operation in $\mathbb{Z}$ is addition and the operation in Z_n is multiplication). Note also that φ is surjective. The fiber of φ over x^a is then

$$\begin{aligned}
\varphi^{-1}(x^a) &= \{m \in \mathbb{Z} \mid x^m = x^a\} \\
&= \{m \in \mathbb{Z} \mid x^{m-a} = 1\} \\
&= \{m \in \mathbb{Z} \mid n \mid m - a\} \qquad \text{(by Proposition 2.3)} \\
&= \{m \in \mathbb{Z} \mid m \equiv a \pmod{n}\} = \bar{a}
\end{aligned}$$

i.e. the fibers of φ are precisely the residue classes modulo n. Figure 1 for this example becomes that shown in Fig. 2. The multiplication in Z_n is just $x^a x^b = x^{a+b}$. The corre-

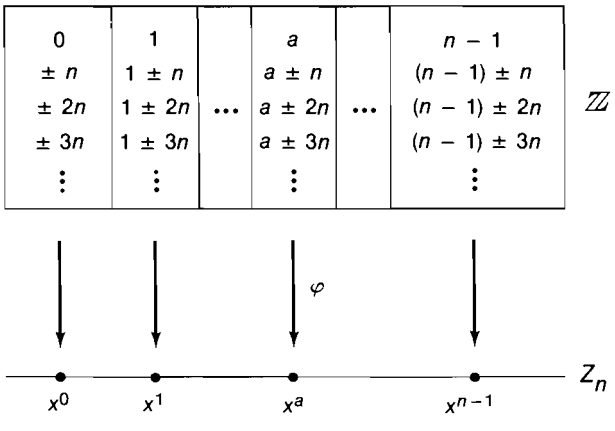


Fig. 2

sponding fibers are $\bar{a}$, b, and $a+b$, so the corresponding group operation for the fibers is $\bar{a} \cdot \bar{b} = \overline{a+b}$. This is just the group $\mathbb{Z}/n\mathbb{Z}$ under addition, a group isomorphic to the image of φ (all of Z_n).

The identity of this group (the fiber above the identity in Z_n) consists of all the multiples of n in $\mathbb{Z}$, namely $n\mathbb{Z}$, a *subgroup* of $\mathbb{Z}$, and the remaining fibers are just translates, $a + n\mathbb{Z}$, of this subgroup. The group operation can also be defined directly by taking *representatives* from these fibers, adding these representatives in $\mathbb{Z}$ and taking the fiber containing this sum (this was the original definition of the group $\mathbb{Z}/n\mathbb{Z}$). From a computational point of view computing the product of $\bar{a}$ and $\bar{b}$ by simply adding representatives a and b is much easier than first computing the image of these fibers under φ (namely, x^a and x^b), multiplying these in H (obtaining x^{a+b}) and then taking the fiber over this product.

We first consider some basic properties of homomorphisms and their fibers. The fiber of a homomorphism $\varphi : G \to H$ lying above the identity of H is given a name:

Definition. If φ is a homomorphism $\varphi : G \to H$, the *kernel* of φ is the set

$$\{g \in G \mid \varphi(g) = 1\}$$

and will be denoted by $\ker \varphi$ (here 1 is the identity of H).

Proposition 1. Let G and H be groups and let $\varphi : G \to H$ be a homomorphism.

(1) $\varphi(1_G) = 1_H$, where 1_G and 1_H are the identities of G and H, respectively.

(2) $\varphi(g^{-1}) = \varphi(g)^{-1}$ for all $g \in G$.

(3) $\varphi(g^n) = \varphi(g)^n$ for all $n \in \mathbb{Z}$.

(4) $\ker \varphi$ is a subgroup of G.

(5) $\operatorname{im}(\varphi)$, the image of G under φ, is a subgroup of H.

Proof: (1) Since $\varphi(1_G) = \varphi(1_G 1_G) = \varphi(1_G)\varphi(1_G)$, the cancellation laws show that (1) holds.

(2) $\varphi(1_G) = \varphi(gg^{-1}) = \varphi(g)\varphi(g^{-1})$ and, by part (1), $\varphi(1_G) = 1_H$, hence

$$1_H = \varphi(g)\varphi(g^{-1}).$$

Multiplying both sides on the left by $\varphi(g)^{-1}$ and simplifying gives (2).

(3) This is an easy exercise in induction for $n \in \mathbb{Z}^+$. By part (2), conclusion (3) holds for negative values of n as well.

(4) Since $1_G \in \ker \varphi$, the kernel of φ is not empty. Let $x, y \in \ker \varphi$, that is $\varphi(x) = \varphi(y) = 1_H$. Then

$$\varphi(xy^{-1}) = \varphi(x)\varphi(y^{-1}) = \varphi(x)\varphi(y)^{-1} = 1_H 1_H^{-1} = 1_H$$

that is, $xy^{-1} \in \ker \varphi$. By the subgroup criterion, $\ker \varphi \leq G$.

(5) Since $\varphi(1_G) = 1_H$, the identity of H lies in the image of φ, so $\operatorname{im}(\varphi)$ is nonempty. If x and y are in $\operatorname{im}(\varphi)$, say $x = \varphi(a)$, $y = \varphi(b)$, then $y^{-1} = \varphi(b^{-1})$ by (2) so that $xy^{-1} = \varphi(a)\varphi(b^{-1}) = \varphi(ab^{-1})$ since φ is a homomorphism. Hence also xy^{-1} is in the image of φ, so $\operatorname{im}(\varphi)$ is a subgroup of H by the subgroup criterion.

We can now define some terminology associated with quotient groups.

Definition. Let $\varphi : G \to H$ be a homomorphism with kernel K. The *quotient group* or *factor group*, G/K (read *G modulo K* or simply *G mod K*), is the group whose elements are the fibers of φ with group operation defined above: namely if X is the fiber above a and Y is the fiber above b then the product of X with Y is defined to be the fiber above the product ab.

The notation emphasizes the fact that the kernel K is a *single element* in the group G/K and we shall see below (Proposition 2) that, as in the case of $\mathbb{Z}/n\mathbb{Z}$ above, the other elements of G/K are just the "translates" of the kernel K. Hence we may think of G/K as being obtained by collapsing or "dividing out" by K (or more precisely, by equivalence modulo K). This explains why G/K is referred to as a "quotient" group.

The definition of the quotient group G/K above requires the map φ explicitly, since the multiplication of the fibers is performed by first projecting the fibers to H via φ, multiplying in H and then determining the fiber over this product. Just as for $\mathbb{Z}/n\mathbb{Z}$ above, it is also possible to define the multiplication of fibers directly in terms of *representatives* from the fibers. This is computationally simpler and the map φ does not enter explicitly. We first show that the fibers of a homomorphism can be expressed in terms of the kernel of the homomorphism just as in the example above (where the kernel was $n\mathbb{Z}$ and the fibers were translates of the form $a + n\mathbb{Z}$).

Proposition 2. Let $\varphi : G \to H$ be a homomorphism of groups with kernel K. Let $X \in G/K$ be the fiber above a, i.e. $X = \varphi^{-1}(a)$. Then

(1) For any $u \in X$, $\quad X = \{uk \mid k \in K\}$

(2) For any $u \in X$, $\quad X = \{ku \mid k \in K\}$.

Proof: We prove (1) and leave the proof of (2) as an exercise. Let $u \in X$ so, by definition of X, $\varphi(u) = a$. Let

$$uK = \{uk \mid k \in K\}.$$

We first prove $uK \subseteq X$. For any $k \in K$,

$$\begin{aligned} \varphi(uk) &= \varphi(u)\varphi(k) && \text{(since } \varphi \text{ is a homomorphism)} \\ &= \varphi(u)1 && \text{(since } k \in \ker \varphi) \\ &= a, \end{aligned}$$

that is, $uk \in X$. This proves $uK \subseteq X$. To establish the reverse inclusion suppose $g \in X$ and let $k = u^{-1}g$. Then

$$\begin{aligned} \varphi(k) &= \varphi(u^{-1})\varphi(g) = \varphi(u)^{-1}\varphi(g) && \text{(by Proposition 1)} \\ &= a^{-1}a = 1. \end{aligned}$$

Thus $k \in \ker \varphi$. Since $k = u^{-1}g$, $g = uk \in uK$, establishing the inclusion $X \subseteq uK$. This proves (1).

The sets arising in Proposition 2 to describe the fibers of a homomorphism φ are defined for *any* subgroup K of G, not necessarily the kernel of some homomorphism

(we shall determine necessary and sufficient conditions for a subgroup to be such a kernel shortly) and are given a name:

Definition. For any $N \leq G$ and any $g \in G$ let

$$gN = \{gn \mid n \in N\} \quad \text{and} \quad Ng = \{ng \mid n \in N\}$$

called respectively a *left coset* and a *right coset* of N in G. Any element of a coset is called a *representative* for the coset.

We have already seen in Proposition 2 that if N is the kernel of a homomorphism and g_1 is any representative for the coset gN then $g_1N = gN$ (and if $g_1 \in Ng$ then $Ng_1 = Ng$). We shall see that this fact is valid for arbitrary subgroups N in Proposition 4 below, which explains the terminology of a *representative*.

If G is an additive group we shall write $g + N$ and $N + g$ for the left and right cosets of N in G with representative g, respectively. In general we can think of the left coset, gN, of N in G as the left translate of N by g. (The reader may wish to review Exercise 18 of Section 1.7 which proves that the right cosets of N in G are precisely the orbits of N acting on G by left multiplication.)

In terms of this definition, Proposition 2 shows that the fibers of a homomorphism are the left cosets of the kernel (and also the right cosets of the kernel), i.e. the elements of the quotient G/K are the left cosets gK, $g \in G$. In the example of $\mathbb{Z}/n\mathbb{Z}$ the multiplication in the quotient group could also be defined in terms of representatives for the cosets. The following result shows the same result is true for G/K in general (provided we know that K is the kernel of some homomorphism), namely that the product of two left cosets X and Y in G/K is computed by choosing any representative u of X, any representative v of Y, multiplying u and v in G and forming the coset $(uv)K$.

Theorem 3. Let G be a group and let K be the kernel of some homomorphism from G to another group. Then the set whose elements are the left cosets of K in G with operation defined by

$$uK \circ vK = (uv)K$$

forms a group, G/K. In particular, this operation is well defined in the sense that if u_1 is any element in uK and v_1 is any element in vK, then $u_1v_1 \in uvK$, i.e. $u_1v_1K = uvK$ so that the multiplication does not depend on the choice of representatives for the cosets. The same statement is true with "right coset" in place of "left coset".

Proof: Let $X, Y \in G/K$ and let $Z = XY$ in G/K, so that by Proposition 2(1) X, Y and Z are (left) cosets of K. By assumption, K is the kernel of some homomorphism $\varphi : G \to H$ so $X = \varphi^{-1}(a)$ and $Y = \varphi^{-1}(b)$ for some $a, b \in H$. By definition of the operation in G/K, $Z = \varphi^{-1}(ab)$. Let u and v be arbitrary representatives of X, Y, respectively, so that $\varphi(u) = a$, $\varphi(v) = b$ and $X = uK$, $Y = vK$. We must show $uv \in Z$. Now

$$\begin{aligned} uv \in Z \quad &\Leftrightarrow \quad uv \in \varphi^{-1}(ab) \\ &\Leftrightarrow \quad \varphi(uv) = ab \\ &\Leftrightarrow \quad \varphi(u)\varphi(v) = ab. \end{aligned}$$

Since the latter equality does hold, $uv \in Z$ hence Z is the (left) coset uvK. (Exercise 2 below shows conversely that every $z \in Z$ can be written as uv, for some $u \in X$ and $v \in Y$.) This proves that the product of X with Y is the coset uvK for any choice of representatives $u \in X$, $v \in Y$ completing the proof of the first statements of the theorem. The last statement in the theorem follows immediately since, by Proposition 2, $uK = Ku$ and $vK = Kv$ for all u and v in G.

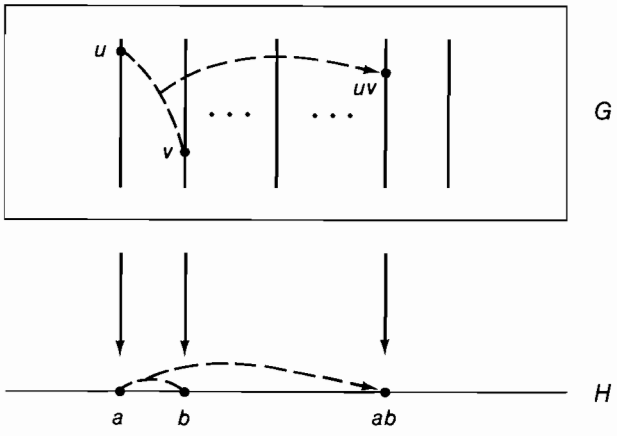


Fig. 3

In terms of Fig. 1, the multiplication in G/K via representatives can be pictured as in Fig. 3.

We emphasize the fact that *the multiplication is independent of the particular representatives chosen*. Namely, the product (or sum, if the group is written additively) of two cosets X and Y is the coset uvK containing the product uv where u and v are *any* representatives for the cosets X and Y, respectively. This process of considering only the coset containing an element, or "reducing mod K" is the same as what we have been doing, in particular, in $\mathbb{Z}/n\mathbb{Z}$. A useful notation for denoting the coset uK containing a representative u is $\bar{u}$. With this notation (which we introduced in the Preliminaries in dealing with $\mathbb{Z}/n\mathbb{Z}$), the quotient group G/K is denoted $\overline{G}$ and the product of elements $\bar{u}$ and $\bar{v}$ is simply the coset containing uv, i.e. $\overline{uv}$. This notation also reinforces the fact that the cosets uK in G/K are *elements* $\bar{u}$ in G/K .

Examples

(1) The first example in this chapter of the homomorphism φ from $\mathbb{Z}$ to Z_n has fibers the left (and also the right) cosets $a + n\mathbb{Z}$ of the kernel $n\mathbb{Z}$. Theorem 3 proves that these cosets form a group under addition of representatives, namely $\mathbb{Z}/n\mathbb{Z}$, which explains the notation for this group. The group is naturally isomorphic to its image under φ, so we recover the isomorphism $\mathbb{Z}/n\mathbb{Z} \cong Z_n$ of Chapter 2.

(2) If $\varphi : G \to H$ is an *isomorphism,* then $K = 1$, the fibers of φ are the singleton subsets of G and so $G/1 \cong G$.

(3) Let G be any group, let $H = 1$ be the group of order 1 and define $\varphi : G \to H$ by $\varphi(g) = 1$, for all $g \in G$. It is immediate that φ is a homomorphism. This map is called the *trivial homomorphism* . Note that in this case $\ker \varphi = G$ and G/G is a group with the single element, G, i.e. $G/G \cong Z_1 = \{1\}$.

(4) Let $G = \mathbb{R}^2$ (operation vector addition), let $H = \mathbb{R}$ (operation addition) and define $\varphi : \mathbb{R}^2 \to \mathbb{R}$ by $\varphi((x, y)) = x$. Thus φ is projection onto the x-axis. We show φ is a homomorphism:

$$\begin{aligned}\varphi((x_1, y_1) + (x_2, y_2)) &= \varphi((x_1 + x_2, y_1 + y_2)) \\ &= x_1 + x_2 = \varphi((x_1, y_1)) + \varphi((x_2, y_2)).\end{aligned}$$

Now

$$\begin{aligned}\ker \varphi &= \{(x, y) \mid \varphi((x, y)) = 0\} \\ &= \{(x, y) \mid x = 0\} = \text{the } y\text{-axis}.\end{aligned}$$

Note that $\ker \varphi$ is indeed a subgroup of $\mathbb{R}^2$ and that the fiber of φ over $a \in \mathbb{R}$ is the translate of the y-axis by a, i.e. the line $x = a$. This is also the left (and the right) coset of the kernel with representative $(a, 0)$ (or any other representative point projecting to a):

$$\overline{(a, 0)} = (a, 0) + \ y\text{-axis}.$$

Hence Fig. 1 in this example becomes Fig. 4.

The group operation (written additively here) can be described either by using the map φ: the sum of the line $(x = a)$ and the line $(x = b)$ is the line $(x = a + b)$; or directly in terms of coset representatives: the sum of the vertical line containing the point (a, y_1) and the vertical line containing the point (b, y_2) is the vertical line containing the point $(a + b, y_1 + y_2)$. Note in particular that the choice of representatives of these vertical lines is not important (i.e. the y-coordinates are not important).

(5) (An example where the group G is non-abelian.) Let $G = Q_8$ and let $H = V_4$ be the Klein 4-group (Section 2.5, Example 2). Define $\varphi : Q_8 \to V_4$ by

$$\varphi(\pm 1) = 1, \qquad \varphi(\pm i) = a, \qquad \varphi(\pm j) = b, \qquad \varphi(\pm k) = c.$$

The check that φ is a homomorphism is left as an exercise — relying on symmetry minimizes the work in showing $\varphi(xy) = \varphi(x)\varphi(y)$ for all x and y in Q_8. It is clear

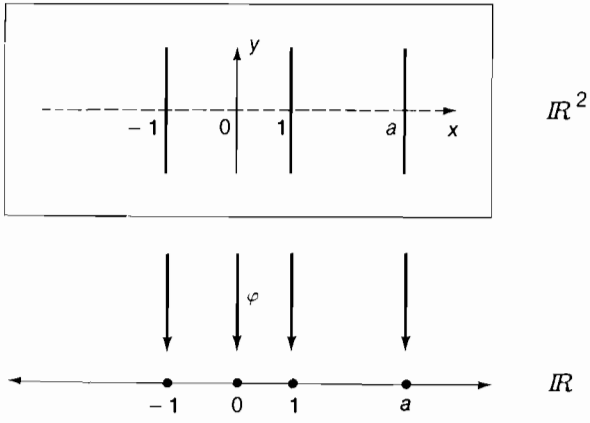


Fig. 4

that φ is surjective and that $\ker \varphi = \{\pm 1\}$. One might think of φ as an "absolute value" function on Q_8 so the fibers of φ are the sets $E = \{\pm 1\}$, $A = \{\pm i\}$, $B = \{\pm j\}$ and $C = \{\pm k\}$, which are collapsed to 1, a, b, and c respectively in $Q_8/\langle \pm 1 \rangle$ and these are the left (and also the right) cosets of $\ker \varphi$ (for example, $A = i \cdot \ker \varphi = \{i, -i\} = \ker \varphi \cdot i$).

By Theorem 3, if we are given a subgroup K of a group G which we know is the kernel of some homomorphism, we may define the quotient G/K without recourse to the homomorphism by the multiplication $uKvK = uvK$. This raises the question whether it is possible to define the quotient group G/N similarly for *any* subgroup N of G. The answer is no in general since this multiplication is not in general well defined (cf. Proposition 5 later). In fact we shall see that it is possible to define the structure of a group on the cosets of N *if and only if* N is the kernel of some homomorphism (Proposition 7). We shall also give a criterion to determine when a subgroup N is such a kernel – this is the notion of a *normal* subgroup and we shall consider non-normal subgroups in subsequent sections.

We first show that the cosets of an arbitrary subgroup of G partition G (i.e. their union is all of G and distinct cosets have trivial intersection).

Proposition 4. Let N be any subgroup of the group G. The set of left cosets of N in G form a partition of G. Furthermore, for all $u, v \in G$, $uN = vN$ if and only if $v^{-1}u \in N$ and in particular, $uN = vN$ if and only if u and v are representatives of the same coset.

Proof: First of all note that since N is a subgroup of G, $1 \in N$. Thus $g = g \cdot 1 \in gN$ for all $g \in G$, i.e.

$$G = \bigcup_{g \in G} gN.$$

To show that distinct left cosets have empty intersection, suppose $uN \cap vN \neq \emptyset$. We show $uN = vN$. Let $x \in uN \cap vN$. Write

$$x = un = vm, \qquad \text{for some } n, m \in N.$$

In the latter equality multiply both sides on the right by n^{-1} to get

$$u = vmn^{-1} = vm_1, \qquad \text{where } m_1 = mn^{-1} \in N.$$

Now for any element ut of uN ($t \in N$),

$$ut = (vm_1)t = v(m_1 t) \in vN.$$

This proves $uN \subseteq vN$. By interchanging the roles of u and v one obtains similarly that $vN \subseteq uN$. Thus two cosets with nonempty intersection coincide.

By the first part of the proposition, $uN = vN$ if and only if $u \in vN$ if and only if $u = vn$, for some $n \in N$ if and only if $v^{-1}u \in N$, as claimed. Finally, $v \in uN$ is equivalent to saying v is a representative for uN, hence $uN = vN$ if and only if u and v are representatives for the same coset (namely the coset $uN = vN$).

Proposition 5. Let G be a group and let N be a subgroup of G.

(1) The operation on the set of left cosets of N in G described by

$$uN \cdot vN = (uv)N$$

is well defined if and only if $gng^{-1} \in N$ for all $g \in G$ and all $n \in N$.

(2) If the above operation is well defined, then it makes the set of left cosets of N in G into a group. In particular the identity of this group is the coset $1N$ and the inverse of gN is the coset $g^{-1}N$ i.e. $(gN)^{-1} = g^{-1}N$.

Proof: (1) Assume first that this operation is well defined, that is, for all $u, v \in G$,

$$\text{if } u, u_1 \in uN \text{ and } v, v_1 \in vN \qquad \text{then} \qquad uvN = u_1v_1N.$$

Let g be an arbitrary element of G and let n be an arbitrary element of N. Letting $u = 1$, $u_1 = n$ and $v = v_1 = g^{-1}$ and applying the assumption above we deduce that

$$1g^{-1}N = ng^{-1}N \quad \text{i.e.} \quad g^{-1}N = ng^{-1}N.$$

Since $1 \in N$, $ng^{-1} \cdot 1 \in ng^{-1}N$. Thus $ng^{-1} \in g^{-1}N$, hence $ng^{-1} = g^{-1}n_1$, for some $n_1 \in N$. Multiplying both sides on the left by g gives $gng^{-1} = n_1 \in N$, as claimed.

Conversely, assume $gng^{-1} \in N$ for all $g \in G$ and all $n \in N$. To prove the operation stated above is well defined let $u, u_1 \in uN$ and $v, v_1 \in vN$. We may write

$$u_1 = un \text{ and } v_1 = vm, \quad \text{for some } n, m \in N.$$

We must prove that $u_1v_1 \in uvN$:

$$\begin{aligned} u_1v_1 &= (un)(vm) = u(vv^{-1})nvm \\ &= (uv)(v^{-1}nv)m = (uv)(n_1m), \end{aligned}$$

where $n_1 = v^{-1}nv = (v^{-1})n(v^{-1})^{-1}$ is an element of N by assumption. Now N is closed under products, so $n_1m \in N$. Thus

$$u_1v_1 = (uv)n_2, \qquad \text{for some } n_2 \in N.$$

Thus the left cosets uvN and u_1v_1N contain the common element u_1v_1. By the preceding proposition they are equal. This proves that the operation is well defined.

(2) If the operation on cosets is well defined the group axioms are easy to check and are induced by their validity in G. For example, the associative law holds because for all $u, v, w \in G$,

$$\begin{aligned} (uN)(vNwN) &= uN(vwN) \\ &= u(vw)N \\ &= (uv)wN = (uNvN)(wN), \end{aligned}$$

since $u(vw) = (uv)w$ in G. The identity in G/N is the coset $1N$ and the inverse of gN is $g^{-1}N$ as is immediate from the definition of the multiplication.

As indicated before, the subgroups N satisfying the condition in Proposition 5 for which there is a natural group structure on the quotient G/N are given a name:

Definition. The element gng^{-1} is called the *conjugate* of $n \in N$ by g. The set $gNg^{-1} = \{gng^{-1} \mid n \in N\}$ is called the *conjugate* of N by g. The element g is said to *normalize* N if $gNg^{-1} = N$. A subgroup N of a group G is called *normal* if every element of G normalizes N, i.e. if $gNg^{-1} = N$ for all $g \in G$. If N is a normal subgroup of G we shall write $N \trianglelefteq G$.

Note that the structure of G is reflected in the structure of the quotient G/N when N is a normal subgroup (for example, the associativity of the multiplication in G/N is induced from the associativity in G and inverses in G/N are induced from inverses in G). We shall see more of the relationship of G to its quotient G/N when we consider the Isomorphism Theorems later in Section 3.

We summarize our results above as Theorem 6.

Theorem 6. Let N be a subgroup of the group G. The following are equivalent:

(1) $N \trianglelefteq G$

(2) $N_G(N) = G$ (recall $N_G(N)$ is the normalizer in G of N)

(3) $gN = Ng$ for all $g \in G$

(4) the operation on left cosets of N in G described in Proposition 5 makes the set of left cosets into a group.

Proof: We have already done the hard equivalences; the others are left as exercises.

As a practical matter, one tries to minimize the computations necessary to determine whether a given subgroup N is normal in a group G. In particular, one tries to avoid as much as possible the computation of all the conjugates gng^{-1} for $n \in N$ and $g \in G$. For example, the elements of N itself normalize N since N is a subgroup. Also, if one has a set of *generators* for N, it suffices to check that all conjugates of these generators lie in N to prove that N is a normal subgroup (this is because the conjugate of a product is the product of the conjugates and the conjugate of the inverse is the inverse of the conjugate)—this is Exercise 26 later. Similarly, if generators for G are also known, then it suffices to check that these generators for G normalize N. In particular, if generators for *both* N and G are known, this reduces the calculations to a small number of conjugations to check. If N is a *finite* group then it suffices to check that the conjugates of a set of generators for N by a set of generators for G are again elements of N (Exercise 29). Finally, it is often possible to prove directly that $N_G(N) = G$ without excessive computations (some examples appear in the next section), again proving that N is a normal subgroup of G without mindlessly computing all possible conjugates gng^{-1}.

We now prove that the normal subgroups are precisely the same as the kernels of homomorphisms considered earlier.

Proposition 7. A subgroup N of the group G is normal if and only if it is the kernel of some homomorphism.

Proof: If N is the kernel of the homomorphism φ, then Proposition 2 shows that the left cosets of N are the same as the right cosets of N (and both are the fibers of the

map φ). By (3) of Theorem 6, N is then a normal subgroup. (Another direct proof of this from the definition of normality for N is given in the exercises).

Conversely, if $N \trianglelefteq G$, let $H = G/N$ and define $\pi : G \to G/N$ by

$$\pi(g) = gN \qquad \text{for all } g \in G.$$

By definition of the operation in G/N,

$$\pi(g_1 g_2) = (g_1 g_2)N = g_1 N g_2 N = \pi(g_1)\pi(g_2).$$

This proves π is a homomorphism. Now

$$\begin{aligned} \ker \pi &= \{g \in G \mid \pi(g) = 1N\} \\ &= \{g \in G \mid gN = 1N\} \\ &= \{g \in G \mid g \in N\} = N. \end{aligned}$$

Thus N is the kernel of the homomorphism π.

The homomorphism π constructed above demonstrating the normal subgroup N as the kernel of a homomorphism is given a name:

Definition. Let $N \trianglelefteq G$. The homomorphism $\pi : G \to G/N$ defined by $\pi(g) = gN$ is called the *natural projection (homomorphism)*[2] of G onto G/N. If $\overline{H} \leq G/N$ is a subgroup of G/N, the *complete preimage* of $\overline{H}$ in G is the preimage of $\overline{H}$ under the natural projection homomorphism.

The complete preimage of a subgroup of G/N is a subgroup of G (cf. Exercise 1 below) which contains the subgroup N since these are the elements which map to the identity $\bar{1} \in \overline{H}$. We shall see in the Isomorphism Theorems in Section 3 that there is a natural correspondence between the subgroups of G that contain N and the subgroups of the quotient G/N.

We now have an "internal" criterion which determines precisely when a subgroup N of a given group G is the kernel of some homomorphism, namely,

$$N_G(N) = G.$$

We may thus think of the normalizer of a subgroup N of G as being a measure of "how close" N is to being a normal subgroup (this explains the choice of name for this subgroup). Keep in mind that the property of being normal is an *embedding* property, that is, it depends on the relation of N to G, not on the internal structure of N itself (the same group N may be a normal subgroup of G but not be normal in a larger group containing G).

We began the discussion of quotient groups with the existence of a homomorphism φ of G to H and showed the kernel of this homomorphism is a normal subgroup N of G and the quotient G/N (defined in terms of fibers originally) is naturally isomorphic to the image of G under φ in H. Conversely, if $N \trianglelefteq G$, we can find a group

[2] The word "natural" has a precise mathematical meaning in the theory of categories; for our purposes we use the term to indicate that the definition of this homomorphism is a "coordinate free" projection, i.e. is described only in terms of the elements themselves, not in terms of generators for G nor N.

H (namely, G/N) and a homomorphism $\pi : G \to H$ such that $\ker \pi = N$ (namely, the natural projection). The study of homomorphic images of G (i.e. the images of homomorphisms from G into other groups) is thus equivalent to the study of quotient groups of G and we shall use homomorphisms to produce normal subgroups and vice versa.

We developed the theory of quotient groups by way of homomorphisms rather than simply defining the notion of a normal subgroup and its associated quotient group to emphasize the fact that the *elements* of the quotient are *subsets* (the fibers or cosets of the kernel N) of the original group G. The visualization in Fig. 1 also emphasizes that N (and its cosets) are projected (or collapsed) onto single elements in the quotient G/N. Computations in the quotient group G/N are performed by taking *representatives* from the various cosets involved.

Some examples of normal subgroups and their associated quotients follow.

Examples

Let G be a group.

(1) The subgroups 1 and G are always normal in G; $G/1 \cong G$ and $G/G \cong 1$.

(2) If G is an *abelian* group, *any* subgroup N of G is normal because for all $g \in G$ and all $n \in N$,

$$gng^{-1} = gg^{-1}n = n \in N.$$

Note that it is important that G be abelian, not just that N be abelian. The structure of G/N may vary as we take different subgroups N of G. For instance, if $G = \mathbb{Z}$, then every subgroup N of G is cyclic:

$$N = \langle n \rangle = \langle -n \rangle = n\mathbb{Z}, \qquad \text{for some } n \in \mathbb{Z}$$

and $G/N = \mathbb{Z}/n\mathbb{Z}$ is a cyclic group with generator $\bar{1} = 1 + n\mathbb{Z}$ (note that 1 is a generator for G).

Suppose now that $G = Z_k$ is the cyclic group of order k. Let x be a generator of G and let $N \leq G$. By Proposition 2.6 $N = \langle x^d \rangle$, where d is the smallest power of x which lies in N. Now

$$G/N = \{gN \mid g \in G\} = \{x^\alpha N \mid \alpha \in \mathbb{Z}\}$$

and since $x^\alpha N = (xN)^\alpha$ (see Exercise 4 below), it follows that

$$G/N = \langle xN \rangle \qquad \text{i.e. } G/N \text{ is cyclic with } xN \text{ as a generator.}$$

By Exercise 5 below, the order of xN in G/N equals d. By Proposition 2.5, $d = \dfrac{|G|}{|N|}$. In summary,

quotient groups of a cyclic group are cyclic

and the image of a generator g for G is a generator $\bar{g}$ for the quotient. If in addition G is a *finite* cyclic group and $N \leq G$, then $|G/N| = \dfrac{|G|}{|N|}$ gives a formula for the order of the quotient group.

(3) If $N \leq Z(G)$, then $N \trianglelefteq G$ because for all $g \in G$ and all $n \in N$, $gng^{-1} = n \in N$, generalizing the previous example (where the center $Z(G)$ is all of G). Thus, in

particular, $Z(G) \trianglelefteq G$. The subgroup $\langle -1 \rangle$ of Q_8 was previously seen to be the kernel of a homomorphism but since $\langle -1 \rangle = Z(Q_8)$ we obtain normality of this subgroup now in another fashion. We already saw that $Q_8/\langle -1 \rangle \cong V_4$. The discussion for D_8 in the next paragraph could be applied equally well to Q_8 to give an independent identification of the isomorphism type of the quotient.

Let $G = D_8$ and let $Z = \langle r^2 \rangle = Z(D_8)$. Since $Z = \{1, r^2\}$, each coset, gZ, consists of the two element set: $\{g, gr^2\}$. Since these cosets partition the 8 elements of D_8 into pairs, there must be 4 (disjoint) left cosets of Z in D_8:

$$\bar{1} = 1Z, \quad \bar{r} = rZ, \quad \bar{s} = sZ, \quad \text{and} \quad \overline{rs} = rsZ.$$

Now by the classification of groups of order 4 (Exercise 10, Section 2.5) we know that $D_8/Z(D_8) \cong Z_4$ or V_4. To determine which of these two is correct (i.e. determine the isomorphism type of the quotient) simply observe that

$$\begin{aligned}(\bar{r})^2 &= r^2 Z = 1Z = \bar{1}\\ (\bar{s})^2 &= s^2 Z = 1Z = \bar{1}\\ (\overline{rs})^2 &= (rs)^2 Z = 1Z = \bar{1}\end{aligned}$$

so every nonidentity element in D_8/Z has order 2. In particular there is no element of order 4 in the quotient, hence D_8/Z is not cyclic so $D_8/Z(D_8) \cong V_4$.

EXERCISES

Let G and H be groups.

1. Let $\varphi : G \to H$ be a homomorphism and let E be a subgroup of H. Prove that $\varphi^{-1}(E) \leq G$ (in words, the *preimage or pullback of a subgroup under a homomorphism is a subgroup*).
2. Let $\varphi : G \to H$ be a homomorphism of groups with kernel K and let $a, b \in \varphi(G)$. Let $X \in G/K$ be the fiber above a and let Y be the fiber above b, i.e. $X = \varphi^{-1}(a)$, $Y = \varphi^{-1}(b)$. Fix an element u of X (so $\varphi(u) = a$). Prove that if $XY = Z$ in the quotient group G/K and w is any member of Z, then there is some $v \in Y$ such that $uv = w$. [Show $u^{-1}w \in Y$.]
3. Let A be an abelian group and let B be a subgroup of A. Prove that A/B is abelian. Give an example of a non-abelian group G containing a proper normal subgroup N such that G/N is abelian.
4. Prove that in the quotient group G/N, $(gN)^\alpha = g^\alpha N$ for all $\alpha \in \mathbb{Z}$.
5. Use the preceding exercise to prove that the order of the element gN in G/N is n, where n is the smallest positive integer such that $g^n \in N$ (and gN has infinite order if no such positive integer exists). Give an example to show that the order of gN in G/N may be strictly smaller than the order of g in G.
6. Define $\varphi : \mathbb{R}^\times \to \{\pm 1\}$ by letting $\varphi(x)$ be x divided by the absolute value of x. Describe the fibers of φ and prove that φ is a homomorphism.
7. Define $\pi : \mathbb{R}^2 \to \mathbb{R}$ by $\pi((x, y)) = x + y$. Prove that π is a surjective homomorphism and describe the kernel and fibers of π geometrically.
8. Let $\varphi : \mathbb{R}^\times \to \mathbb{R}^\times$ be the map sending x to the absolute value of x. Prove that φ is a homomorphism and find the image of φ. Describe the kernel and the fibers of φ.
9. Define $\varphi : \mathbb{C}^\times \to \mathbb{R}^\times$ by $\varphi(a + bi) = a^2 + b^2$. Prove that φ is a homomorphism and find the image of φ. Describe the kernel and the fibers of φ geometrically (as subsets of the plane).

10. Let $\varphi : \mathbb{Z}/8\mathbb{Z} \to \mathbb{Z}/4\mathbb{Z}$ by $\varphi(\bar{a}) = \bar{a}$. Show that this is a well defined, surjective homomorphism and describe its fibers and kernel explicitly (showing that φ is well defined involves the fact that $\bar{a}$ has a different meaning in the domain and range of φ).

11. Let F be a field and let $G = \{\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in F,\ ac \neq 0\} \leq GL_2(F)$.

    (a) Prove that the map $\varphi : \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mapsto a$ is a surjective homomorphism from G onto $F^\times$ (recall that $F^\times$ is the multiplicative group of nonzero elements in F). Describe the fibers and kernel of φ.

    (b) Prove that the map $\psi : \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mapsto (a, c)$ is a surjective homomorphism from G onto $F^\times \times F^\times$. Describe the fibers and kernel of ψ.

    (c) Let $H = \{\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \mid b \in F\}$. Prove that H is isomorphic to the additive group F.

12. Let G be the additive group of real numbers, let H be the multiplicative group of complex numbers of absolute value 1 (the unit circle in the complex plane) and let $\varphi : G \to H$ be the homomorphism $\varphi : r \mapsto e^{2\pi i r}$. Draw the points on a real line which lie in the kernel of φ. Describe similarly the elements in the fibers of φ above the points $-1, i$, and $e^{4\pi i/3}$ of H. (Figure 1 of the text for this homomorphism φ is usually depicted using the diagram in Fig. 5.)

13. Repeat the preceding exercise with the map φ replaced by the map $\varphi : r \mapsto e^{4\pi i r}$.

14. Consider the additive quotient group $\mathbb{Q}/\mathbb{Z}$.

    (a) Show that every coset of $\mathbb{Z}$ in $\mathbb{Q}$ contains exactly one representative $q \in \mathbb{Q}$ in the range $0 \leq q < 1$.

    (b) Show that every element of $\mathbb{Q}/\mathbb{Z}$ has finite order but that there are elements of arbitrarily large order.

    (c) Show that $\mathbb{Q}/\mathbb{Z}$ is the torsion subgroup of $\mathbb{R}/\mathbb{Z}$ (cf. Exercise 6, Section 2.1).

    (d) Prove that $\mathbb{Q}/\mathbb{Z}$ is isomorphic to the multiplicative group of roots of unity in $\mathbb{C}^\times$.

15. Prove that a quotient of a divisible abelian group by any proper subgroup is also divisible. Deduce that $\mathbb{Q}/\mathbb{Z}$ is divisible (cf. Exercise 19, Section 2.4).

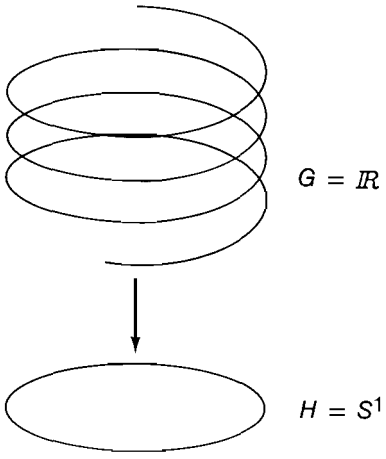


Fig. 5

16. Let G be a group, let N be a normal subgroup of G and let $\overline{G} = G/N$. Prove that if $G = \langle x, y\rangle$ then $\overline{G} = \langle \overline{x}, \overline{y}\rangle$. Prove more generally that if $G = \langle S\rangle$ for any subset S of G, then $\overline{G} = \langle \overline{S}\rangle$.

17. Let G be the dihedral group of order 16 (whose lattice appears in Section 2.5):

$$G = \langle r, s \mid r^8 = s^2 = 1,\ rs = sr^{-1}\rangle$$

and let $\overline{G} = G/\langle r^4\rangle$ be the quotient of G by the subgroup generated by r^4 (this subgroup is the center of G, hence is normal).

(a) Show that the order of $\overline{G}$ is 8.

(b) Exhibit each element of $\overline{G}$ in the form $\overline{s}^a\overline{r}^b$, for some integers a and b.

(c) Find the order of each of the elements of $\overline{G}$ exhibited in (b).

(d) Write each of the following elements of $\overline{G}$ in the form $\overline{s}^a\overline{r}^b$, for some integers a and b as in (b): $\overline{rs}$, $\overline{sr^{-2}s}$, $\overline{s^{-1}r^{-1}sr}$.

(e) Prove that $\overline{H} = \langle \overline{s}, \overline{r}^2\rangle$ is a normal subgroup of $\overline{G}$ and $\overline{H}$ is isomorphic to the Klein 4-group. Describe the isomorphism type of the complete preimage of $\overline{H}$ in G.

(f) Find the center of $\overline{G}$ and describe the isomorphism type of $\overline{G}/Z(\overline{G})$.

18. Let G be the quasidihedral group of order 16 (whose lattice was computed in Exercise 11 of Section 2.5):

$$G = \langle \sigma, \tau \mid \sigma^8 = \tau^2 = 1,\ \sigma\tau = \tau\sigma^3\rangle$$

and let $\overline{G} = G/\langle \sigma^4\rangle$ be the quotient of G by the subgroup generated by σ^4 (this subgroup is the center of G, hence is normal).

(a) Show that the order of $\overline{G}$ is 8.

(b) Exhibit each element of $\overline{G}$ in the form $\overline{\tau}^a\overline{\sigma}^b$, for some integers a and b.

(c) Find the order of each of the elements of $\overline{G}$ exhibited in (b).

(d) Write each of the following elements of $\overline{G}$ in the form $\overline{\tau}^a\overline{\sigma}^b$, for some integers a and b as in (b): $\overline{\sigma\tau}$, $\overline{\tau\sigma^{-2}\tau}$, $\overline{\tau^{-1}\sigma^{-1}\tau\sigma}$.

(e) Prove that $\overline{G} \cong D_8$.

19. Let G be the modular group of order 16 (whose lattice was computed in Exercise 14 of Section 2.5):

$$G = \langle u, v \mid u^2 = v^8 = 1,\ vu = uv^5\rangle$$

and let $\overline{G} = G/\langle v^4\rangle$ be the quotient of G by the subgroup generated by v^4 (this subgroup is contained in the center of G, hence is normal).

(a) Show that the order of $\overline{G}$ is 8.

(b) Exhibit each element of $\overline{G}$ in the form $\overline{u}^a\overline{v}^b$, for some integers a and b.

(c) Find the order of each of the elements of $\overline{G}$ exhibited in (b).

(d) Write each of the following elements of $\overline{G}$ in the form $\overline{u}^a\overline{v}^b$, for some integers a and b as in (b): $\overline{vu}$, $\overline{uv^{-2}u}$, $\overline{u^{-1}v^{-1}uv}$.

(e) Prove that $\overline{G}$ is abelian and is isomorphic to $Z_2 \times Z_4$.

20. Let $G = \mathbb{Z}/24\mathbb{Z}$ and let $\widetilde{G} = G/\langle \overline{12}\rangle$, where for each integer a we simplify notation by writing $\widetilde{\overline{a}}$ as $\widetilde{a}$.

(a) Show that $\widetilde{G} = \{\widetilde{0}, \widetilde{1}, \ldots, \widetilde{11}\}$.

(b) Find the order of each element of $\widetilde{G}$.

(c) Prove that $\widetilde{G} \cong \mathbb{Z}/12\mathbb{Z}$. (Thus $(\mathbb{Z}/24\mathbb{Z})/(12\mathbb{Z}/24\mathbb{Z}) \cong \mathbb{Z}/12\mathbb{Z}$, just as if we inverted and cancelled the $24\mathbb{Z}$'s.)

21. Let $G = Z_4 \times Z_4$ be given in terms of the following generators and relations:

$$G = \langle x, y \mid x^4 = y^4 = 1,\ xy = yx \rangle.$$

Let $\overline{G} = G/\langle x^2 y^2 \rangle$ (note that every subgroup of the abelian group G is normal).

(a) Show that the order of $\overline{G}$ is 8.

(b) Exhibit each element of $\overline{G}$ in the form $\overline{x}^a \overline{y}^b$, for some integers a and b.

(c) Find the order of each of the elements of $\overline{G}$ exhibited in (b).

(d) Prove that $\overline{G} \cong Z_4 \times Z_2$.

22. **(a)** Prove that if H and K are normal subgroups of a group G then their intersection $H \cap K$ is also a normal subgroup of G.

(b) Prove that the intersection of an arbitrary nonempty collection of normal subgroups of a group is a normal subgroup (do not assume the collection is countable).

23. Prove that the join (cf. Section 2.5) of any nonempty collection of normal subgroups of a group is a normal subgroup.

24. Prove that if $N \trianglelefteq G$ and H is any subgroup of G then $N \cap H \trianglelefteq H$.

25. Prove that a subgroup N of a group G is normal if and only if $gNg^{-1} \subseteq N$ for all $g \in G$. (For an arbitrary subgroup N the set $\{g \in G \mid gNg^{-1} \subseteq N\}$ need not be closed under inverses—we shall see an example of this in Section 5.5—so do not try to prove the more general statement that this set equals the normalizer of N in G).

26. Let $a, b \in G$.

(a) Prove that the conjugate of the product of a and b is the product of the conjugate of a and the conjugate of b. Prove that the order of a and the order of any conjugate of a are the same.

(b) Prove that the conjugate of a^{-1} is the inverse of the conjugate of a.

(c) Let $N = \langle S \rangle$ for some subset S of G. Prove that $N \trianglelefteq G$ if $gSg^{-1} \subseteq N$ for all $g \in G$.

(d) Deduce that if N is the cyclic group $\langle x \rangle$, then N is normal in G if and only if for each $g \in G$, $gxg^{-1} = x^k$ for some $k \in \mathbb{Z}$.

(e) Let n be a positive integer. Prove that the subgroup N of G generated by all the elements of G of order n is a normal subgroup of G.

27. Let N be a *finite* subgroup of a group G. Show that $gNg^{-1} \subseteq N$ if and only if $gNg^{-1} = N$. Deduce that $N_G(N) = \{g \in G \mid gNg^{-1} \subseteq N\}$.

28. Let N be a *finite* subgroup of a group G and assume $N = \langle S \rangle$ for some subset S of G. Prove that an element $g \in G$ normalizes N if and only if $gSg^{-1} \subseteq N$.

29. Let N be a *finite* subgroup of G and suppose $G = \langle T \rangle$ and $N = \langle S \rangle$ for some subsets S and T of G. Prove that N is normal in G if and only if $tSt^{-1} \subseteq N$ for all $t \in T$.

30. Let $N \leq G$ and let $g \in G$. Prove that $gN = Ng$ if and only if $g \in N_G(N)$.

31. Prove that if $H \leq G$ and N is a normal subgroup of H then $H \leq N_G(N)$. Deduce that $N_G(N)$ is the largest subgroup of G in which N is normal (i.e. is the join of all subgroups H for which $N \trianglelefteq H$).

32. Prove that every subgroup of Q_8 is normal. For each subgroup find the isomorphism type of its corresponding quotient. [You may use the lattice of subgroups for Q_8 in Section 2.5.]

33. Find all normal subgroups of D_8 and for each of these find the isomorphism type of its corresponding quotient. [You may use the lattice of subgroups for D_8 in Section 2.5.]

34. Let $D_{2n} = \langle r, s \mid r^n = s^2 = 1,\ rs = sr^{-1} \rangle$ be the usual presentation of the dihedral group of order $2n$ and let k be a positive integer dividing n.

(a) Prove that $\langle r^k \rangle$ is a normal subgroup of D_{2n}.

(b) Prove that $D_{2n}/\langle r^k \rangle \cong D_{2k}$.

35. Prove that $SL_n(F) \trianglelefteq GL_n(F)$ and describe the isomorphism type of the quotient group (cf. Exercise 9, Section 2.1).

36. Prove that if $G/Z(G)$ is cyclic then G is abelian. [If $G/Z(G)$ is cyclic with generator $xZ(G)$, show that every element of G can be written in the form $x^a z$ for some integer $a \in \mathbb{Z}$ and some element $z \in Z(G)$.]

37. Let A and B be groups. Show that $\{(a, 1) \mid a \in A\}$ is a normal subgroup of $A \times B$ and the quotient of $A \times B$ by this subgroup is isomorphic to B.

38. Let A be an abelian group and let D be the (diagonal) subgroup $\{(a, a) \mid a \in A\}$ of $A \times A$. Prove that D is a normal subgroup of $A \times A$ and $(A \times A)/D \cong A$.

39. Suppose A is the non-abelian group S_3 and D is the diagonal subgroup $\{(a, a) \mid a \in A\}$ of $A \times A$. Prove that D is not normal in $A \times A$.

40. Let G be a group, let N be a normal subgroup of G and let $\overline{G} = G/N$. Prove that $\overline{x}$ and $\overline{y}$ commute in $\overline{G}$ if and only if $x^{-1}y^{-1}xy \in N$. (The element $x^{-1}y^{-1}xy$ is called the *commutator* of x and y and is denoted by $[x, y]$.)

41. Let G be a group. Prove that $N = \langle x^{-1}y^{-1}xy \mid x, y \in G\rangle$ is a normal subgroup of G and G/N is abelian (N is called the *commutator subgroup* of G).

42. Assume both H and K are normal subgroups of G with $H \cap K = 1$. Prove that $xy = yx$ for all $x \in H$ and $y \in K$. [Show $x^{-1}y^{-1}xy \in H \cap K$.]

43. Assume $\mathcal{P} = \{A_i \mid i \in I\}$ is any partition of G with the property that $\mathcal{P}$ is a group under the "quotient operation" defined as follows: to compute the product of A_i with A_j take any element a_i of A_i and any element a_j of A_j and let A_iA_j be the element of $\mathcal{P}$ containing a_ia_j (this operation is assumed to be well defined). Prove that the element of $\mathcal{P}$ that contains the identity of G is a normal subgroup of G and the elements of $\mathcal{P}$ are the cosets of this subgroup (so $\mathcal{P}$ is just a quotient group of G in the usual sense).

3.2 MORE ON COSETS AND LAGRANGE'S THEOREM

In this section we continue the study of quotient groups. Since for finite groups one of the most important invariants of a group is its order we first prove that the order of a quotient group of a finite group can be readily computed: $|G/N| = \dfrac{|G|}{|N|}$. In fact we derive this as a consequence of a more general result, Lagrange's Theorem (see Exercise 19, Section 1.7). This theorem is one of the most important combinatorial results in finite group theory and will be used repeatedly. After indicating some easy consequences of Lagrange's Theorem we study more subtle questions concerning cosets of non-normal subgroups.

The proof of Lagrange's Theorem is straightforward and important. It is the same line of reasoning we used in Example 3 of the preceding section to compute $|D_8/Z(D_8)|$.

Theorem 8. (Lagrange) If G is a finite group and H is a subgroup of G, then the order of H divides the order of G (i.e. $|H| \mid |G|$) and the number of left cosets of H in G equals $\dfrac{|G|}{|H|}$.

Proof: Let $|H| = n$ and let the number of left cosets of H in G equal k. By Proposition 4 the set of left cosets of H in G partition G. By definition of a left coset

the map:

$$H \to gH \qquad \text{defined by} \qquad h \mapsto gh$$

is a surjection from H to the left coset gH. The left cancellation law implies this map is injective since $gh_1 = gh_2$ implies $h_1 = h_2$. This proves that H and gH have the same order:

$$|gH| = |H| = n.$$

Since G is partitioned into k disjoint subsets each of which has cardinality n, $|G| = kn$. Thus $k = \dfrac{|G|}{n} = \dfrac{|G|}{|H|}$, completing the proof.

Definition. If G is a group (possibly infinite) and $H \leq G$, the number of left cosets of H in G is called the *index* of H in G and is denoted by $|G : H|$.

In the case of finite groups the index of H in G is $\dfrac{|G|}{|H|}$. For G an infinite group the quotient $\dfrac{|G|}{|H|}$ does not make sense. Infinite groups may have subgroups of finite or infinite index (e.g. $\{0\}$ is of infinite index in $\mathbb{Z}$ and $\langle n \rangle$ is of index n in $\mathbb{Z}$ for every $n > 0$).

We now derive some easy consequences of Lagrange's Theorem.

Corollary 9. If G is a finite group and $x \in G$, then the order of x divides the order of G. In particular $x^{|G|} = 1$ for all x in G.

Proof: By Proposition 2.2, $|x| = |\langle x \rangle|$. The first part of the corollary follows from Lagrange's Theorem applied to $H = \langle x \rangle$. The second statement is clear since now $|G|$ is a multiple of the order of x.

Corollary 10. If G is a group of prime order p, then G is cyclic, hence $G \cong Z_p$.

Proof: Let $x \in G$, $x \neq 1$. Thus $|\langle x \rangle| > 1$ and $|\langle x \rangle|$ divides $|G|$. Since $|G|$ is prime we must have $|\langle x \rangle| = |G|$, hence $G = \langle x \rangle$ is cyclic (with any nonidentity element x as generator). Theorem 2.4 completes the proof.

With Lagrange's Theorem in hand we examine some additional examples of normal subgroups.

Examples

(1) Let $H = \langle (1\ 2\ 3) \rangle \leq S_3$ and let $G = S_3$. We show $H \trianglelefteq S_3$. As noted in Section 2.2,

$$H \leq N_G(H) \leq G.$$

By Lagrange's Theorem, the order of H divides the order of $N_G(H)$ and the order of $N_G(H)$ divides the order of G. Since G has order 6 and H has order 3, the only possibilities for $N_G(H)$ are H or G. A direct computation gives

$$(1\ 2)(1\ 2\ 3)(1\ 2) = (1\ 3\ 2) = (1\ 2\ 3)^{-1}.$$

Since $(1\ 2) = (1\ 2)^{-1}$, this calculation shows that $(1\ 2)$ conjugates a generator of H to another generator of H. By Exercise 24 of Section 2.3 this is sufficient to prove that $(1\ 2) \in N_G(H)$. Thus $N_G(H) \neq H$ so $N_G(H) = G$, i.e. $H \trianglelefteq S_3$, as claimed. This argument illustrates that checking normality of a subgroup can often be reduced to a small number of calculations. A generalization of this example is given in the next example.

(2) Let G be any group containing a subgroup H of index 2. We prove $H \trianglelefteq G$. Let $g \in G - H$ so, by hypothesis, the two left cosets of H in G are $1H$ and gH. Since $1H = H$ and the cosets partition G, we must have $gH = G - H$. Now the two right cosets of H in G are $H1$ and Hg. Since $H1 = H$, we again must have $Hg = G - H$. Combining these gives $gH = Hg$, so every left coset of H in G is a right coset. By Theorem 6, $H \trianglelefteq G$. By definition of index, $|G/H| = 2$, so that $G/H \cong Z_2$. One must be careful to appreciate that the reason H is normal in this case is not because we can choose the same coset representatives 1 and g for both the left and right cosets of H but that there is a type of pidgeon-hole principle at work: since $1H = H = H1$ for any subgroup H of any group G, the index assumption forces the remaining elements to comprise the remaining coset (either left or right). We shall see that this result is itself a special case of a result we shall prove in the next chapter.

Note that this result proves that $\langle i \rangle$, $\langle j \rangle$ and $\langle k \rangle$ are normal subgroups of Q_8 and that $\langle s, r^2 \rangle$, $\langle r \rangle$ and $\langle sr, r^2 \rangle$ are normal subgroups of D_8.

(3) The property "is a normal subgroup of" is not transitive. For example,

$$\langle s \rangle \trianglelefteq \langle s, r^2 \rangle \trianglelefteq D_8$$

(each subgroup is of index 2 in the next), however, $\langle s \rangle$ is not normal in D_8 because $rsr^{-1} = sr^2 \notin \langle s \rangle$.

We now examine some examples of non-normal subgroups. Although in abelian groups every subgroup is normal, this is not the case in non-abelian groups (in some sense Q_8 is the unique exception to this). In fact, there are groups G in which the only normal subgroups are the trivial ones: 1 and G. Such groups are called *simple groups* (simple does not mean easy, however). Simple groups play an important role in the study of general groups and this role will be described in Section 4. For now we emphasize that not every subgroup of a group G is normal in G, indeed normal subgroups may be quite rare in G. The search for normal subgroups of a given group is in general a highly nontrivial problem.

Examples

(1) Let $H = \langle (1\ 2) \rangle \leq S_3$. Since H is of prime index 3 in S_3, by Lagrange's Theorem the only possibilities for $N_{S_3}(H)$ are H or S_3. Direct computation shows:

$$(1\ 3)(1\ 2)(1\ 3)^{-1} = (1\ 3)(1\ 2)(1\ 3) = (2\ 3) \notin H$$

so $N_{S_3}(H) \neq S_3$, that is, H is not a normal subgroup of S_3. One can also see this by considering the left and right cosets of H; for instance

$$(1\ 3)H = \{(1\ 3), (1\ 2\ 3)\} \qquad \text{and} \qquad H(1\ 3) = \{(1\ 3), (1\ 3\ 2)\}.$$

Since the left coset $(1\ 3)H$ is the unique left coset of H containing $(1\ 3)$, the right coset $H(1\ 3)$ cannot be a left coset (see also Exercise 6). Note also that the "group operation" on the left cosets of H in S_3 defined by multiplying representatives is

not even well defined. For example, consider the product of the two left cosets $1H$ and $(1\,3)H$. The elements 1 and $(1\,2)$ are both representatives for the coset $1H$, yet $1\cdot(1\,3) = (1\,3)$ and $(1\,2)\cdot(1\,3) = (1\,3\,2)$ are not both elements of the same left coset as they should be if the product of these cosets were independent of the particular representatives chosen. This is an example of Theorem 6 which states that the cosets of a subgroup form a group *only* when the subgroup is a normal subgroup.

(2) Let $G = S_n$ for some $n \in \mathbb{Z}^+$ and fix some $i \in \{1, 2, ..., n\}$. As in Section 2.2 let

$$G_i = \{\sigma \in G \mid \sigma(i) = i\}$$

be the stabilizer of the point i. Suppose $\tau \in G$ and $\tau(i) = j$. It follows directly from the definition of G_i that for all $\sigma \in G_i$, $\tau\sigma(i) = j$. Furthermore, if $\mu \in G$ and $\mu(i) = j$, then $\tau^{-1}\mu(i) = i$, that is, $\tau^{-1}\mu \in G_i$, so $\mu \in \tau G_i$. This proves that

$$\tau G_i = \{\mu \in G \mid \mu(i) = j\},$$

i.e. the left coset τG_i consists of the permutations in S_n which take i to j. We can clearly see that distinct left cosets have empty intersection and that the number of distinct left cosets equals the number of distinct images of the integer i under the action of G, namely there are n distinct left cosets. Thus $|G : G_i| = n$. Using the same notation and by similar reasoning we see that

$$G_i\tau = \{\lambda \in G \mid \lambda(j) = i\},$$

i.e. the right coset $G_i\tau$ consists of the permutations in S_n which take j to i. If $n > 2$, for any nonidentity element τ, $\tau G_i \neq G_i\tau$ since there are certainly permutations which take i to j but do not take j to i, so G_i is not a normal subgroup. In fact $N_G(G_i) = G_i$ by Exercise 30 of Section 1, so G_i is in some sense far from being normal in S_n. This example generalizes the preceding one.

(3) In D_8 the only subgroup of order 2 which is normal is the center: $\langle r^2 \rangle$.

We shall see many more examples of non-normal subgroups as we develop the theory.

The *full converse* to Lagrange's Theorem is *not* true: namely, if G is a finite group and n divides $|G|$, then G need not have a subgroup of order n. For example, let A be the group of symmetries of a regular tetrahedron. By Exercise 9 of Section 1.2, $|A| = 12$. Suppose A had a subgroup H of order 6. Since $\dfrac{|A|}{|H|} = 2$, H would be of index 2 in A, hence $H \trianglelefteq A$ and $A/H \cong Z_2$. Since the quotient group has order 2, the square of every element in the quotient is the identity, so for all $g \in A$, $(gH)^2 = 1H$, that is, for all $g \in A$, $g^2 \in H$. If g is an element of A of order 3, we obtain $g = (g^2)^2 \in H$, that is, H must contain all elements of A of order 3. This is a contradiction since $|H| = 6$ but one can easily exhibit 8 rotations of a tetrahedron of order 3.

There are some partial converses to Lagrange's Theorem. For finite *abelian* groups the full converse of Lagrange is true, namely an abelian group has a subgroup of order n for each divisor n of $|G|$ (in fact, this holds under weaker assumptions than "abelian"; we shall see this in Chapter 6). A partial converse which holds for arbitrary finite groups is:

Theorem 11. (Cauchy) If G is a finite group and p is a prime dividing $|G|$, then G has an element of order p.

Proof: We shall give a proof of this in the next chapter and another elegant proof is outlined in Exercise 9.

The strongest converse to Lagrange's Theorem which applies to *arbitrary* finite groups is:

Theorem 12. (Sylow) If G is a finite group of order $p^\alpha m$, where p is a prime and p does not divide m, then G has a subgroup of order p^α.

We shall prove this theorem in the next chapter and derive more information on the number of subgroups of order p^α.

We conclude this section with some useful results involving cosets.

Definition. Let H and K be subgroups of a group and define $HK = \{hk \mid h \in H,\ k \in K\}$.

Proposition 13. If H and K are finite subgroups of a group,

$$|HK| = \frac{|H||K|}{|H \cap K|}.$$

Proof: Notice that HK is a union of left cosets of K, namely,

$$HK = \bigcup_{h \in H} hK.$$

Since each coset of K has $|K|$ elements it suffices to find the number of *distinct* left cosets of the form hK, $h \in H$. But $h_1K = h_2K$ for $h_1, h_2 \in H$ if and only if $h_2^{-1}h_1 \in K$. Thus

$$h_1K = h_2K \quad \Leftrightarrow \quad h_2^{-1}h_1 \in H \cap K \quad \Leftrightarrow \quad h_1(H \cap K) = h_2(H \cap K).$$

Thus the number of distinct cosets of the form hK, for $h \in H$ is the number of distinct cosets $h(H \cap K)$, for $h \in H$. The latter number, by Lagrange's Theorem, equals $\dfrac{|H|}{|H \cap K|}$. Thus HK consists of $\dfrac{|H|}{|H \cap K|}$ distinct cosets of K (each of which has $|K|$ elements) which gives the formula above.

Notice that there was no assumption that HK be a subgroup in Proposition 13. For example, if $G = S_3$, $H = \langle(1\,2)\rangle$ and $K = \langle(2\,3)\rangle$, then $|H| = |K| = 2$ and $|H \cap K| = 1$, so $|HK| = 4$. By Lagrange's Theorem HK cannot be a subgroup. As a consequence, we must have $S_3 = \langle(1\,2), (2\,3)\rangle$.

Proposition 14. If H and K are subgroups of a group, HK is a subgroup if and only if $HK = KH$.

Proof: Assume first that $HK = KH$ and let $a, b \in HK$. We prove $ab^{-1} \in HK$ so HK is a subgroup by the subgroup criterion. Let

$$a = h_1k_1 \qquad \text{and} \qquad b = h_2k_2,$$

for some $h_1, h_2 \in H$ and $k_1, k_2 \in K$. Thus $b^{-1} = k_2^{-1}h_2^{-1}$, so $ab^{-1} = h_1k_1k_2^{-1}h_2^{-1}$. Let $k_3 = k_1k_2^{-1} \in K$ and $h_3 = h_2^{-1}$. Thus $ab^{-1} = h_1k_3h_3$. Since $HK = KH$,

$$k_3h_3 = h_4k_4, \qquad \text{for some } h_4 \in H, \quad k_4 \in K.$$

Thus $ab^{-1} = h_1h_4k_4$, and since $h_1h_4 \in H$, $k_4 \in K$, we obtain $ab^{-1} \in HK$, as desired.

Conversely, assume that HK is a subgroup of G. Since $K \leq HK$ and $H \leq HK$, by the closure property of subgroups, $KH \subseteq HK$. To show the reverse containment let $hk \in HK$. Since HK is assumed to be a subgroup, write $hk = a^{-1}$, for some $a \in HK$. If $a = h_1k_1$, then

$$hk = (h_1k_1)^{-1} = k_1^{-1}h_1^{-1} \in KH,$$

completing the proof.

Note that $HK = KH$ does *not* imply that the elements of H commute with those of K (contrary to what the notation may suggest) but rather that every product hk is of the form $k'h'$ (h need not be h' nor k be k') and conversely. For example, if $G = D_{2n}$, $H = \langle r \rangle$ and $K = \langle s \rangle$, then $G = HK = KH$ so that HK is a subgroup and $rs = sr^{-1}$ so the elements of H do not commute with the elements of K. This is an example of the following sufficient condition for HK to be a subgroup:

Corollary 15. If H and K are subgroups of G and $H \leq N_G(K)$, then HK is a subgroup of G. In particular, if $K \trianglelefteq G$ then $HK \leq G$ for any $H \leq G$.

Proof: We prove $HK = KH$. Let $h \in H$, $k \in K$. By assumption, $hkh^{-1} \in K$, hence

$$hk = (hkh^{-1})h \in KH.$$

This proves $HK \subseteq KH$. Similarly, $kh = h(h^{-1}kh) \in HK$, proving the reverse containment. The corollary follows now from the preceding proposition.

Definition. If A is any subset of $N_G(K)$ (or $C_G(K)$), we shall say A *normalizes* K (*centralizes* K, respectively).

With this terminology, Corollary 15 states that *HK is a subgroup if H normalizes K* (similarly, *HK is a subgroup if K normalizes H*).

In some instances one can prove that a finite group is a product of two of its subgroups by simply using the order formula in Proposition 13. For example, let $G = S_4$, $H = D_8$ and let $K = \langle (1\,2\,3) \rangle$, where we consider D_8 as a subgroup of S_4 by identifying each symmetry with its permutation on the 4 vertices of a square (under some fixed labelling). By Lagrange's Theorem, $H \cap K = 1$ (see Exercise 8). Proposition 13 then shows $|HK| = 24$ hence we must have $HK = S_4$. Since HK is a group, $HK = KH$. We leave as an exercise the verification that neither H nor K normalizes the other (so Corollary 15 could not have been used to give $HK = KH$).

Finally, throughout this chapter we have worked with left cosets of a subgroup. The same combinatorial results could equally well have been proved using right cosets. For normal subgroups this is trivial since left and right cosets are the same, but for

non-normal subgroups some left cosets are not right cosets (for any choice of representative) so some (simple) verifications are necessary. For example, Lagrange's Theorem gives that in a finite group G

$$\textit{the number of right cosets of the subgroup } H \textit{ is } \frac{|G|}{|H|}.$$

Thus in a finite group the *number* of left cosets of H in G equals the *number* of right cosets even though the left cosets are not right cosets in general. This is also true for infinite groups as Exercise 12 below shows. Thus for purely combinatorial purposes one may use either left or right cosets (but not a mixture when a partition of G is needed). Our consistent use of left cosets is somewhat arbitrary although it will have some benefits when we discuss actions on cosets in the next chapter. Readers may encounter in some works the notation $H \setminus G$ to denote the set of right cosets of H in G.

In some papers one may also see the notation G/H used to denote the set of left cosets of H in G even when H is not normal in G (in which case G/H is called the *coset space* of left cosets of H in G). We shall not use this notation.

EXERCISES

Let G be a group.

1. Which of the following are permissible orders for subgroups of a group of order 120: 1, 2, 5, 7, 9, 15, 60, 240? For each permissible order give the corresponding index.
2. Prove that the lattice of subgroups of S_3 in Section 2.5 is correct (i.e. prove that it contains all subgroups of S_3 and that their pairwise joins and intersections are correctly drawn).
3. Prove that the lattice of subgroups of Q_8 in Section 2.5 is correct.
4. Show that if $|G| = pq$ for some primes p and q (not necessarily distinct) then either G is abelian or $Z(G) = 1$. [See Exercise 36 in Section 1.]
5. Let H be a subgroup of G and fix some element $g \in G$.
   (a) Prove that gHg^{-1} is a subgroup of G of the same order as H.
   (b) Deduce that if $n \in \mathbb{Z}^+$ and H is the unique subgroup of G of order n then $H \trianglelefteq G$.
6. Let $H \leq G$ and let $g \in G$. Prove that if the right coset Hg equals *some* left coset of H in G then it equals the left coset gH and g must be in $N_G(H)$.
7. Let $H \leq G$ and define a relation $\sim$ on G by $a \sim b$ if and only if $b^{-1}a \in H$. Prove that $\sim$ is an equivalence relation and describe the equivalence class of each $a \in G$. Use this to prove Proposition 4.
8. Prove that if H and K are finite subgroups of G whose orders are relatively prime then $H \cap K = 1$.
9. This exercise outlines a proof of Cauchy's Theorem due to James McKay [Another Proof of Cauchy's Group Theorem, *Amer. Math. Monthly*, 66(1959), p. 119]. Let G be a finite group and let p be a prime dividing $|G|$. Let $\mathcal{S}$ denote the set of p-tuples of elements of G the product of whose coordinates is 1:

$$\mathcal{S} = \{(x_1, x_2, \ldots, x_p) \mid x_i \in G \text{ and } x_1 x_2 \cdots x_p = 1\}.$$

   (a) Show that $\mathcal{S}$ has $|G|^{p-1}$ elements, hence has order divisible by p.
   Define the relation $\sim$ on $\mathcal{S}$ by letting $\alpha \sim \beta$ if β is a cyclic permutation of α.

(b) Show that a cyclic permutation of an element of S is again an element of S.
(c) Prove that $\sim$ is an equivalence relation on S.
(d) Prove that an equivalence class contains a single element if and only if it is of the form $(x, x, \ldots, x)$ with $x^p = 1$.
(e) Prove that every equivalence class has order 1 or p (this uses the fact that p is a *prime*). Deduce that $|G|^{p-1} = k + pd$, where k is the number of classes of size 1 and d is the number of classes of size p.
(f) Since $\{(1, 1, \ldots, 1)\}$ is an equivalence class of size 1, conclude from (e) that there must be a nonidentity element x in G with $x^p = 1$, i.e. G contains an element of order p. [Show $p \mid k$ and so $k > 1$.]

10. Let H and K be subgroups of finite index in the (possibly infinite) group G with $|G : H| = m$ and $|G : K| = n$. Prove that l.c.m.$(m, n) \leq |G : H \cap K| \leq mn$. Deduce that if m and n are relatively prime then $|G : H \cap K| = |G : H| \cdot |G : K|$.

11. Let $H \leq K \leq G$. Prove that $|G : H| = |G : K| \cdot |K : H|$ (do not assume G is a finite group).

12. Let $H \leq G$. Prove that the map $x \mapsto x^{-1}$ sends each left coset of H in G onto a right coset of H and gives a bijection between the set of left cosets and the set of right cosets of H in G (hence the number of left cosets of H in G equals the number of right cosets).

13. Fix any labelling of the vertices of a square and use this to identify D_8 as a subgroup of S_4. Prove that the elements of D_8 and $\langle(1\,2\,3)\rangle$ do not commute in S_4.

14. Prove that S_4 does not have a normal subgroup of order 8 nor a normal subgroup of order 3.

15. Let $G = S_n$ and for fixed $i \in \{1, 2, \ldots, n\}$ let G_i be the stabilizer of i. Prove that $G_i \cong S_{n-1}$.

16. Use Lagrange's Theorem in the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ to prove *Fermat's Little Theorem*: if p is a prime then $a^p \equiv a (\text{mod } p)$ for all $a \in \mathbb{Z}$.

17. Let p be a prime and let n be a positive integer. Find the order of $\bar{p}$ in $(\mathbb{Z}/(p^n - 1)\mathbb{Z})^\times$ and deduce that $n \mid \varphi(p^n - 1)$ (here φ is Euler's function).

18. Let G be a finite group, let H be a subgroup of G and let $N \trianglelefteq G$. Prove that if $|H|$ and $|G : N|$ are relatively prime then $H \leq N$.

19. Prove that if N is a normal subgroup of the finite group G and $(|N|, |G : N|) = 1$ then N is the unique subgroup of G of order $|N|$.

20. Let A be an abelian normal subgroup of G and let B be any subgroup of G. Prove that $A \cap B \trianglelefteq AB$.

21. Prove that $\mathbb{Q}$ has no proper subgroups of finite index. Deduce that $\mathbb{Q}/\mathbb{Z}$ has no proper subgroups of finite index. [Recall Exercise 21, Section 1.6 and Exercise 15, Section 1.]

22. Use Lagrange's Theorem in the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$ to prove *Euler's Theorem*: $a^{\varphi(n)} \equiv 1 \bmod n$ for every integer a relatively prime to n, where φ denotes Euler's φ–function.

23. Determine the last two digits of $3^{3^{100}}$. [Determine $3^{100} \bmod \varphi(100)$ and use the previous exercise.]

3.3 THE ISOMORPHISM THEOREMS

In this section we derive some straightforward consequences of the relations between quotient groups and homomorphisms which were discussed in Section 1. In particular

we consider the relation between the lattice of subgroups of a quotient group, G/N, and the lattice of subgroups of the group G. The first result restates our observations in Section 1 on the relation of the image of a homomorphism to the quotient by the kernel (sometimes called the Fundamental Theorem of Homomorphisms):

Theorem 16. (The First Isomorphism Theorem) If $\varphi : G \to H$ is a homomorphism of groups, then $\ker \varphi \trianglelefteq G$ and $G/\ker \varphi \cong \varphi(G)$.

Corollary 17. Let $\varphi : G \to H$ be a homomorphism of groups.

(1) φ is injective if and only if $\ker \varphi = 1$.

(2) $|G : \ker \varphi| = |\varphi(G)|$.

Proof: Exercise.

When we consider abstract vector spaces we shall see that Corollary 17(2) gives a formula possibly already familiar from the theory of linear transformations: if $\varphi : V \to W$ is a linear transformation of vector spaces, then $\dim V = \operatorname{rank} \varphi + \operatorname{nullity} \varphi$.

Theorem 18. (The Second or Diamond Isomorphism Theorem) Let G be a group, let A and B subgroups of G and assume $A \leq N_G(B)$. Then $A \cap B \trianglelefteq A$ and $AB/B \cong A/A \cap B$.

Proof: By Corollary 15 AB is a subgroup of G. To show $A \cap B \trianglelefteq A$ let $a \in A$ and let $x \in A \cap B$. Since both elements are in A, $axa^{-1} \in A$. Since A normalizes B, $axa^{-1} \in B$. Together these prove $axa^{-1} \in A \cap B$, as desired.

Define $\varphi : AB \to A/A \cap B$ by $\varphi(ab) = a(A \cap B)$, for $a \in A$ and $b \in B$. We must show φ is well defined since each element of AB may have many representations as a product of the form ab. Suppose $ab = a_1b_1$, for some $a_1 \in A$ and $b_1 \in B$; we must show $\varphi(ab) = \varphi(a_1b_1)$, i.e. $a(A \cap B) = a_1(A \cap B)$. Since $ab = a_1b_1$, $a_1^{-1}a = b_1b^{-1}$ so $a_1^{-1}a \in B$. Since a and a_1 are members of A, $a_1^{-1}a \in A \cap B$. Thus $a(A \cap B) = a_1(A \cap B)$, which proves φ is well defined.

To show φ is a homomorphism let $a, a_1 \in A$ and let $b, b_1 \in B$, so

$$\begin{aligned}
\varphi((ab)(a_1b_1)) &= \varphi(aba_1b_1) \\
&= \varphi(a(a_1a_1^{-1})ba_1b_1) \\
&= \varphi(aa_1(a_1^{-1}ba_1)b_1) \\
&= \varphi((aa_1)(b_2b_1)) && \text{(where } b_2 = a_1^{-1}ba_1 \in B \text{ as } a_1 \in N_G(B)\text{)} \\
&= aa_1(A \cap B) && \text{(by definition of } \varphi) \\
&= a(A \cap B)a_1(A \cap B) && \text{(by the group operation in } A/A \cap B) \\
&= \varphi(ab)\varphi(a_1b_1) && \text{(by definition of } \varphi).
\end{aligned}$$

The map φ is surjective since for any $a(A \cap B) \in A/A \cap B$, the element a belongs to AB and $\varphi(a) = a(A \cap B)$.

Finally,

$$\begin{aligned}
\ker \varphi &= \{ab \mid \varphi(ab) = 1(A \cap B)\} \\
&= \{ab \mid a(A \cap B) = 1(A \cap B)\} \\
&= \{ab \mid a \in A \cap B\}.
\end{aligned}$$

Since $B \subseteq \{ab \mid a \in A \cap B\}$ (because $1 \in A$) and the reverse containment is clear since $A \cap B \leq B$, we have $B = \ker \varphi$. By the First Isomorphism Theorem, $AB/B \cong A/A \cap B$, completing the proof.

Note that this gives a new proof of the order formula in Proposition 13 in the special case that $A \leq N_G(B)$. The reason this theorem is called the Diamond Isomorphism is because of the portion of the lattice of subgroups of G involved (see Fig. 6). The markings in the lattice lines indicate which quotients are isomorphic. The "quotient" AB/A need not be a group (i.e. A need not be normal in AB), however we still have $|AB : A| = |B : A \cap B|$.

The third Isomorphism Theorem considers the question of taking quotient groups of quotient groups.

Theorem 19. (The Third Isomorphism Theorem) Let G be a group and let H and K be normal subgroups of G with $H \leq K$. Then $K/H \trianglelefteq G/H$ and

$$(G/H)/(K/H) \cong G/K.$$

If we denote the quotient by H with a bar, this can be written

$$\overline{G}/\overline{K} \cong G/K.$$

Proof: We leave as an easy exercise the verification that $K/H \trianglelefteq G/H$. Define

$$\begin{aligned} \varphi : G/H &\to G/K \\ (gH) &\mapsto gK. \end{aligned}$$

To show φ is well defined suppose $g_1H = g_2H$. Then $g_1 = g_2h$, for some $h \in H$. Because $H \leq K$, the element h is also an element of K, hence $g_1K = g_2K$ i.e. $\varphi(g_1H) = \varphi(g_2H)$, which shows φ is well defined. Since g may be chosen arbitrarily in G, φ is a surjective homomorphism. Finally,

$$\begin{aligned} \ker \varphi &= \{gH \in G/H \mid \varphi(gH) = 1K\} \\ &= \{gH \in G/H \mid gK = 1K\} \\ &= \{gH \in G/H \mid g \in K\} = K/H. \end{aligned}$$

By the First Isomorphism Theorem, $(G/H)/(K/H) \cong G/K$.

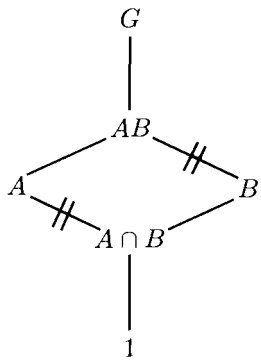


Fig. 6

An easy aid for remembering the Third Isomorphism Theorem is: "invert and cancel" (as one would for fractions). This theorem shows that we gain no new structural information from taking quotients of a quotient group.

The final isomorphism theorem describes the relation between the lattice of subgroups of the quotient group G/N and the lattice of subgroups of G. The lattice for G/N can be read immediately from the lattice for G by collapsing the group N to the identity. More precisely, there is a one-to-one correspondence between the subgroups of G containing N and the subgroups of G/N, so that the lattice for G/N (or rather, an isomorphic copy) appears in the lattice for G as the collection of subgroups of G between N and G. In particular, the lattice for G/N appears at the "top" of the lattice for G, a result we mentioned at the beginning of the chapter.

Theorem 20. (The Fourth or Lattice Isomorphism Theorem) Let G be a group and let N be a normal subgroup of G. Then there is a bijection from the set of subgroups A of G which contain N onto the set of subgroups $\overline{A} = A/N$ of G/N. In particular, every subgroup of $\overline{G}$ is of the form A/N for some subgroup A of G containing N (namely, its preimage in G under the natural projection homomorphism from G to G/N). This bijection has the following properties: for all $A, B \leq G$ with $N \leq A$ and $N \leq B$,

(1) $A \leq B$ if and only if $\overline{A} \leq \overline{B}$,

(2) if $A \leq B$, then $|B : A| = |\overline{B} : \overline{A}|$,

(3) $\overline{\langle A, B\rangle} = \langle \overline{A}, \overline{B}\rangle$,

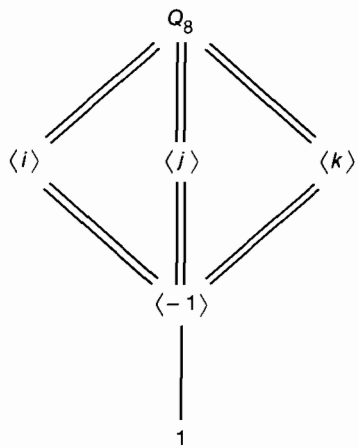
(4) $\overline{A \cap B} = \overline{A} \cap \overline{B}$, and

(5) $A \trianglelefteq G$ if and only if $\overline{A} \trianglelefteq \overline{G}$.

Proof: The complete preimage of a subgroup in G/N is a subgroup of G by Exercise 1 of Section 1. The numerous details of the theorem to check are all completely straightforward. We therefore leave the proof of this theorem to the exercises.

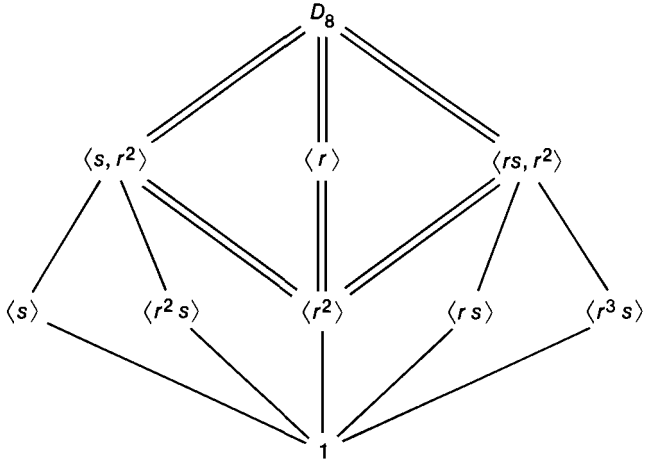
Examples

(1) Let $G = Q_8$ and let N be the normal subgroup $\langle -1\rangle$. The (isomorphic copy of the) lattice of G/N consists of the double lines in the lattice of G below. Note that we



previously proved that $Q_8/\langle -1\rangle \cong V_4$ and the two lattices do indeed coincide (see Section 2.5 for the lattices of Q_8 and V_4).

(2) The same process gives us the lattice of $D_8/\langle r^2\rangle$ (the double lines) in the lattice of D_8:



Note that in the second example above there are subgroups of G which do not directly correspond to subgroups in the quotient group G/N, namely the subgroups of G which do not contain the normal subgroup N. This is because the subgroup N projects to a point in G/N and so several subgroups of G can project to the same subgroup in the quotient. The image of the subgroup H of G under the natural projection homomorphism from G to G/N is the same as the image of the subgroup HN of G, and the subgroup HN of G contains N. Conversely, the preimage of a subgroup $\overline{H}$ of G/N contains N and is the unique subgroup of G containing N whose image in G/N is $\overline{H}$. It is the subgroups of G containing N which appear explicitly in the lattice for G/N.

The two lattices of groups of order 8 above emphasize the fact that the isomorphism type of a group cannot in general be determined from the knowledge of the isomorphism types of G/N and N, since $Q_8/\langle -1\rangle \cong D_8/\langle r^2\rangle$ and $\langle -1\rangle \cong \langle r^2\rangle$ yet Q_8 and D_8 are not isomorphic. We shall discuss this question further in the next section.

We shall often indicate the index of one subgroup in another in the lattice of subgroups, as follows:

$$\begin{array}{c} A \\ | \quad n \\ B \end{array}$$

where the integer n equals $|A : B|$. For example, all the unbroken edges in the lattices of Q_8 and D_8 would be labelled with 2. Thus the order of any subgroup, A, is the product of all integers which label any path upward from the identity to A. Also, by Theorem 20(2) these indices remain unchanged in quotients of G by normal subgroups of G contained in B, i.e. the portion of the lattice for G corresponding to the lattice of the quotient group has the correct indices for the quotient as well.

Finally we include a remark concerning the definition of homomorphisms on quotient groups. We have, in the course of the proof of the isomorphism theorems,

$$\begin{tikzcd} G \arrow[r, "\pi"] \arrow[dr, "\Phi"'] & G/N \arrow[d, "\varphi"] \arrow[dl, phantom, "\circlearrowright"] \\ & H \end{tikzcd}$$

Fig. 7

encountered situations where a homomorphism φ on the quotient group G/N is specified by giving the value of φ on the coset gN in terms of the representative g alone. In each instance we then had to prove φ was well defined, i.e. was independent of the choice of g. In effect we are defining a homomorphism, Φ, on G itself by specifying the value of φ at g. Then independence of g is equivalent to requiring that Φ be trivial on N, so that

$$\varphi \textit{ is well defined on } G/N \textit{ if and only if } N \leq \ker \Phi.$$

This gives a simple criterion for defining homomorphisms on quotients (namely, define a homomorphism on G and check that N is contained in its kernel). In this situation we shall say the homomorphism Φ *factors through* N and φ is the *induced* homomorphism on G/N. This can be denoted pictorially as in Fig. 7, where the circular arrow indicates that $\Phi = \varphi \circ \pi$, i.e. the image in H of an element in G does not depend on which path one takes in the diagram above. If this is the case, then the diagram is said to *commute*.

At this point we have developed all the background material so that Section 6.3 on free groups and presentations may now be read.

EXERCISES

Let G be a group.

1. Let F be a finite field of order q and let n be a positive integer. Prove that $|GL_n(F) : SL_n(F)| = q - 1$. [See Exercise 35, Section 1.]
2. Prove all parts of the Lattice Isomorphism Theorem.
3. Prove that if H is a normal subgroup of G of prime index p then for all $K \leq G$ either:
   (i) $K \leq H$ or
   (ii) $G = HK$ and $|K : K \cap H| = p$.
4. Let C be a normal subgroup of the group A and let D be a normal subgroup of the group B. Prove that $(C \times D) \trianglelefteq (A \times B)$ and $(A \times B)/(C \times D) \cong (A/C) \times (B/D)$.
5. Let $QD_{16} = \langle \sigma, \tau \rangle$ be the quasidihedral group described in Exercise 11 of Section 2.5. Prove that $\langle \sigma^4 \rangle$ is normal in QD_{16} and use the Lattice Isomorphism Theorem to draw the lattice of subgroups of $QD_{16}/\langle \sigma^4 \rangle$. Which group of order 8 has the same lattice as this quotient? Use generators and relations for $QD_{16}/\langle \sigma^4 \rangle$ to decide the isomorphism type of this group.
6. Let $M = \langle v, u \rangle$ be the modular group of order 16 described in Exercise 14 of Section 2.5. Prove that $\langle v^4 \rangle$ is normal in M and use the Lattice Isomorphism Theorem to draw the lattice of subgroups of $M/\langle v^4 \rangle$. Which group of order 8 has the same lattice as this quotient? Use generators and relations for $M/\langle v^4 \rangle$ to decide the isomorphism type of this group.
7. Let M and N be normal subgroups of G such that $G = MN$. Prove that $G/(M \cap N) \cong (G/M) \times (G/N)$. [Draw the lattice.]
8. Let p be a prime and let G be the group of p-power roots of 1 in $\mathbb{C}$ (cf. Exercise 18, Sec-

tion 2.4). Prove that the map $z \mapsto z^p$ is a surjective homomorphism. Deduce that G is isomorphic to a proper quotient of itself.

9. Let p be a prime and let G be a group of order $p^a m$, where p does not divide m. Assume P is a subgroup of G of order p^a and N is a normal subgroup of G of order $p^b n$, where p does not divide n. Prove that $|P \cap N| = p^b$ and $|PN/N| = p^{a-b}$. (The subgroup P of G is called a *Sylow p-subgroup* of G. This exercise shows that the intersection of any Sylow p-subgroup of G with a normal subgroup N is a Sylow p-subgroup of N.)

10. Generalize the preceding exercise as follows. A subgroup H of a finite group G is called a *Hall subgroup* of G if its index in G is relatively prime to its order: $(|G : H|, |H|) = 1$. Prove that if H is a Hall subgroup of G and $N \trianglelefteq G$, then $H \cap N$ is a Hall subgroup of N and HN/N is a Hall subgroup of G/N.

3.4 COMPOSITION SERIES AND THE HÖLDER PROGRAM

The remarks in the preceding section on lattices leave us with the intuitive picture that a quotient group G/N is the group whose structure (e.g. lattice) describes the structure of G "above" the normal subgroup N. Although this is somewhat vague, it gives at least some notion of the driving force behind one of the most powerful techniques in finite group theory (and even some branches of infinite group theory): the use of induction. In many instances the application of an inductive procedure follows a pattern similar to the following proof of a special case of Cauchy's Theorem. Although Cauchy's Theorem is valid for arbitrary groups (cf. Exercise 9 of Section 2 above), the following is a good example of the use of information on a normal subgroup N and on the quotient G/N to determine information about G, and we shall need this particular result in Chapter 4.

Proposition 21. If G is a finite abelian group and p is a prime dividing $|G|$, then G contains an element of order p.

Proof: The proof proceeds by induction on $|G|$, namely, we assume the result is valid for every group whose order is strictly smaller than the order of G and then prove the result valid for G (this is sometimes referred to as *complete* induction). Since $|G| > 1$, there is an element $x \in G$ with $x \neq 1$. If $|G| = p$ then x has order p by Lagrange's Theorem and we are done. We may therefore assume $|G| > p$.

Suppose p divides $|x|$ and write $|x| = pn$. By Proposition 2.5(3), $|x^n| = p$, and again we have an element of order p. We may therefore assume p does not divide $|x|$.

Let $N = \langle x \rangle$. Since G is abelian, $N \trianglelefteq G$. By Lagrange's Theorem,

$$|G/N| = \frac{|G|}{|N|}$$

and since $N \neq 1$, $|G/N| < |G|$. Since p does not divide $|N|$, we must have $p \mid |G/N|$. We can now apply the induction assumption to the smaller group G/N to conclude it contains an element, $\bar{y} = yN$, of order p. Since $y \notin N$ ($\bar{y} \neq \bar{1}$) but $y^p \in N$ ($\bar{y}^p = \bar{1}$), we must have $\langle y^p \rangle \neq \langle y \rangle$, that is, $|y^p| < |y|$. Proposition 2.5(2) implies $p \mid |y|$. We are now in the situation described in the preceding paragraph, so that argument again produces an element of order p. The induction is complete.

The philosophy behind this method of proof is that if we have a sufficient amount of information about some normal subgroup, N, of a group G and sufficient information on G/N, then somehow we can piece this information together to force G itself to have some desired property. The induction comes into play because both N and G/N have smaller order than G. In general, just how much data is required is a delicate matter since, as we have already seen, the full isomorphism type of G cannot be determined from the isomorphism types of N and G/N alone.

Clearly a basic obstruction to this approach is the necessity of producing a normal subgroup, N, of G with $N \neq 1$ or G. In the preceding argument this was easy since G was abelian. Groups with no nontrivial proper normal subgroups are fundamental obstructions to this method of proof.

Definition. A (finite or infinite) group G is called *simple* if $|G| > 1$ and the only normal subgroups of G are 1 and G.

By Lagrange's Theorem if $|G|$ is a prime, its only subgroups (let alone normal ones) are 1 and G, so G is simple. In fact, every abelian simple group is isomorphic to Z_p, for some prime p (cf. Exercise 1). There are non-abelian simple groups (of both finite and infinite order), the smallest of which has order 60 (we shall introduce this group as a member of an infinite family of simple groups in the next section).

Simple groups, by definition, cannot be "factored" into pieces like N and G/N and as a result they play a role analogous to that of the primes in the arithmetic of $\mathbb{Z}$. This analogy is supported by a "unique factorization theorem" (for finite groups) which we now describe.

Definition. In a group G a sequence of subgroups

$$1 = N_0 \leq N_1 \leq N_2 \leq \cdots \leq N_{k-1} \leq N_k = G$$

is called a *composition series* if $N_i \trianglelefteq N_{i+1}$ and N_{i+1}/N_i a simple group, $0 \leq i \leq k-1$. If the above sequence is a composition series, the quotient groups N_{i+1}/N_i are called *composition factors* of G.

Keep in mind that it is not assumed that each $N_i \trianglelefteq G$, only that $N_i \trianglelefteq N_{i+1}$. Thus

$$1 \trianglelefteq \langle s \rangle \trianglelefteq \langle s, r^2 \rangle \trianglelefteq D_8 \quad \text{and} \quad 1 \trianglelefteq \langle r^2 \rangle \trianglelefteq \langle r \rangle \trianglelefteq D_8$$

are two composition series for D_8 and in each series there are 3 composition factors, each of which is isomorphic to (the simple group) Z_2.

Theorem 22. (Jordan–Hölder) Let G be a finite group with $G \neq 1$. Then

(1) G has a composition series and

(2) The composition factors in a composition series are unique, namely, if $1 = N_0 \leq N_1 \leq \cdots \leq N_r = G$ and $1 = M_0 \leq M_1 \leq \cdots \leq M_s = G$ are two composition series for G, then $r = s$ and there is some permutation, π, of $\{1, 2, \ldots, r\}$ such that

$$M_{\pi(i)+1}/M_{\pi(i)} \cong N_{i+1}/N_i, \qquad 1 \leq i \leq r-1.$$

Proof: This is fairly straightforward. Since we shall not explicitly use this theorem to prove others in the text we outline the proof in a series of exercises at the end of this section.

Thus every finite group has a "factorization" (i.e. composition series) and although the series itself need not be unique (as D_8 shows) the number of composition factors and their isomorphism types are uniquely determined. Furthermore, nonisomorphic groups may have the same (up to isomorphism) list of composition factors (see Exercise 2). This motivates a two-part program for classifying all finite groups up to isomorphism:

The Hölder Program:

(1) classify all finite simple groups, and

(2) find all ways of "putting simple groups together" to form other groups.

These two problems form part of an underlying motivation for much of the development of group theory. Analogues of these problems may also be found as recurring themes throughout mathematics. We include a few more comments on the current status of progress on these problems.

The classification of finite simple groups (part (1) of the Hölder Program) was completed in 1980, about 100 years after the formulation of the Hölder Program. Efforts by over 100 mathematicians covering between 5,000 and 10,000 journal pages (spread over some 300 to 500 individual papers) proves the following:

Theorem. There is a list consisting of 18 (infinite) families of simple groups and 26 simple groups not belonging to these families (the *sporadic* simple groups) such that every finite simple group is isomorphic to one of the groups in this list.

One example of a family of simple groups is $\{Z_p \mid p \text{ a prime}\}$. A second infinite family in the list of finite simple groups is:

$$\{SL_n(\mathbb{F})/Z(SL_n(\mathbb{F})) \mid n \in \mathbb{Z}^+, n \geq 2 \text{ and } \mathbb{F} \text{ a finite field }\}.$$

These groups are all simple except for $SL_2(\mathbb{F}_2)$ and $SL_2(\mathbb{F}_3)$ where $\mathbb{F}_2$ is the finite field with 2 elements and $\mathbb{F}_3$ is the finite field with 3 elements. This is a 2-parameter family (n and $\mathbb{F}$ being independent parameters). We shall not prove these groups are simple (although it is not technically beyond the scope of the text) but rather refer the reader to [*Finite Group Theory* by M. Aschbacher, Cambridge University Press, 1986] for proofs and an extensive discussion of the simple group problem. A third family of finite simple groups, the alternating groups, is discussed in the next section; we shall prove these groups are simple in the next chapter.

To gain some idea of the complexity of the classification of finite simple groups the reader may wish to peruse the proof of one of the cornerstones of the entire classification:

Theorem. (Feit–Thompson) If G is a simple group of odd order, then $G \cong Z_p$ for some prime p.

This proof takes 255 pages[3] of hard mathematics.

Part (2) of the Hölder Program, sometimes called the *extension problem*, was rather vaguely formulated. A more precise description of "putting two groups together" is: given groups A and B, describe how to obtain all groups G containing a normal subgroup N such that $N \cong B$ and $G/N \cong A$. For instance, if $A = B = Z_2$, there are precisely two possibilities for G, namely, Z_4 and V_4 (see Exercise 10 of Section 2.5) and the Hölder program seeks to describe how the two groups of order 4 could have been built from two Z_2's without a priori knowledge of the existence of the groups of order 4. This part of the Hölder Program is extremely difficult, even when the subgroups involved are of small order. For example, all composition factors of a group G have order 2 if and only if $|G| = 2^n$, for some n (one implication is easy and we shall prove both implications in Chapter 6). It is known, however, that the number of nonisomorphic groups of order 2^n grows (exponentially) as a function of 2^n, so the number of ways of putting groups of 2-power order together is not bounded. Nonetheless, there are a wealth of interesting and powerful techniques in this subtle area which serve to unravel the structure of large classes of groups. We shall only discuss a couple of ways of building larger groups from smaller ones (in the sense above) but even from this limited excursion into the area of group extensions we shall construct numerous new examples of groups and prove some classification theorems.

One class of groups which figures prominently in the theory of polynomial equations is the class of *solvable* groups:

Definition. A group G is *solvable* if there is a chain of subgroups

$$1 = G_0 \trianglelefteq G_1 \trianglelefteq G_2 \trianglelefteq \ldots \trianglelefteq G_s = G$$

such that G_{i+1}/G_i is abelian for $i = 0, 1, \ldots, s-1$.

The terminology comes from the correspondence in Galois Theory between these groups and polynomials which can be solved by radicals (which essentially means there is an algebraic formula for the roots). Exercise 5 shows that finite solvable groups are precisely those groups whose composition factors are all of prime order.

One remarkable property of finite solvable groups is the following generalization of Sylow's Theorem due to Phillip Hall:

Theorem. The finite group G is solvable if and only if for every divisor n of $|G|$ such that $(n, \frac{|G|}{n}) = 1$, G has a subgroup of order n.

As another illustration of how properties of a group G can be deduced from combined information from a normal subgroup N and the quotient group G/N we prove:

$$\textit{if } N \textit{ and } G/N \textit{ are solvable, then } G \textit{ is solvable.}$$

To see this let $\overline{G} = G/N$, let $1 = N_0 \trianglelefteq N_1 \trianglelefteq \ldots \trianglelefteq N_n = N$ be a chain of subgroups of N such that N_{i+1}/N_i is abelian, $0 \leq i < n$ and let $\overline{1} = \overline{G_0} \trianglelefteq \overline{G_1} \trianglelefteq \ldots \trianglelefteq \overline{G_m} = \overline{G}$ be a chain of subgroups of $\overline{G}$ such that $\overline{G_{i+1}}/\overline{G_i}$ is abelian, $0 \leq i < m$. By the Lattice

[3] The solvability of groups of odd order, *Pacific Journal of Mathematics*, 13(1963).

Isomorphism Theorem there are subgroups G_i of G with $N \leq G_i$ such that $G_i/N = \overline{G_i}$ and $G_i \trianglelefteq G_{i+1}$, $0 \leq i < m$. By the Third Isomorphism Theorem

$$\overline{G_{i+1}}/\overline{G_i} = (G_{i+1}/N)/(G_i/N) \cong G_{i+1}/G_i.$$

Thus

$$1 = N_0 \trianglelefteq N_1 \trianglelefteq \ldots \trianglelefteq N_n = N = G_0 \trianglelefteq G_1 \trianglelefteq \ldots \trianglelefteq G_m = G$$

is a chain of subgroups of G all of whose successive quotient groups are abelian. This proves G is solvable.

It is inaccurate to say that finite group theory is concerned *only* with the Hölder Program. It *is* accurate to say that the Hölder Program suggests a large number of problems and motivates a number of algebraic techniques. For example, in the study of the extension problem where we are given groups A and B and wish to find G and $N \trianglelefteq G$ with $N \cong B$ and $G/N \cong A$, we shall see that (under certain conditions) we are led to an *action* of the group A on the set B. Such actions form the crux of the next chapter (and will result in information both about simple and non-simple groups) and this notion is a powerful one in mathematics not restricted to the theory of groups.

The final section of this chapter introduces another family of groups and although in line with our interest in simple groups, will be of independent importance throughout the text, particularly in our study later of determinants and the solvability of polynomial equations.

EXERCISES

1. Prove that if G is an abelian simple group then $G \cong Z_p$ for some prime p (do not assume G is a finite group).
2. Exhibit all 3 composition series for Q_8 and all 7 composition series for D_8. List the composition factors in each case.
3. Find a composition series for the quasidihedral group of order 16 (cf. Exercise 11, Section 2.5). Deduce that QD_{16} is solvable.
4. Use Cauchy's Theorem and induction to show that a finite abelian group has a subgroup of order n for each positive divisor n of its order.
5. Let G be a *finite* group. Prove that the following are equivalent:
   (i) G is solvable
   (ii) G has a chain of subgroups: $1 = H_0 \trianglelefteq H_1 \trianglelefteq H_2 \trianglelefteq \ldots \trianglelefteq H_s = G$ such that H_{i+1}/H_i is cyclic, $0 \leq i \leq s-1$
   (iii) all composition factors of G are of prime order
   (iv) G has a chain of subgroups: $1 = N_0 \trianglelefteq N_1 \trianglelefteq N_2 \trianglelefteq \ldots \trianglelefteq N_t = G$ such that each N_i is a normal subgroup of G and N_{i+1}/N_i is abelian, $0 \leq i \leq t-1$.
6. Prove part (1) of the Jordan–Hölder Theorem by induction on $|G|$.
7. If G is a finite group and $H \trianglelefteq G$ prove that there is a composition series of G, one of whose terms is H.
8. Prove that subgroups and quotient groups of a solvable group are solvable.
9. Prove the following special case of part (2) of the Jordan–Hölder Theorem: assume the finite group G has two composition series

$$1 = N_0 \trianglelefteq N_1 \trianglelefteq \ldots \trianglelefteq N_r = G \qquad \text{and} \qquad 1 = M_0 \trianglelefteq M_1 \trianglelefteq M_2 = G.$$

Show that $r = 2$ and that the list of composition factors is the same. [Use the Second Isomorphism Theorem.]

10. Prove part (2) of the Jordan–Hölder Theorem by induction on $\min\{r, s\}$. [Apply the inductive hypothesis to $H = N_{r-1} \cap M_{s-1}$ and use the preceding exercises.]

11. Prove that if H is a nontrivial normal subgroup of the solvable group G then there is a nontrivial subgroup A of H with $A \trianglelefteq G$ and A abelian.

12. Prove (without using the Feit–Thompson Theorem) that the following are equivalent:
(i) every group of odd order is solvable
(ii) the only simple groups of odd order are those of prime order.

3.5 TRANSPOSITIONS AND THE ALTERNATING GROUP

Transpositions and Generation of S_n

As we saw in Section 1.3 (and will prove in the next chapter) every element of S_n can be written as a product of disjoint cycles in an essentially unique fashion. In contrast, every element of S_n can be written in many different ways as a (nondisjoint) product of cycles. For example, even in S_3 the element $\sigma = (1\,2\,3)$ may be written

$$\sigma = (1\,2\,3) = (1\,3)(1\,2) = (1\,2)(1\,3)(1\,2)(1\,3) = (1\,2)(2\,3)$$

and, in fact, there are an infinite number of different ways to write σ. Not requiring the cycles to be disjoint totally destroys the uniqueness of a representation of a permutation as a product of cycles. We can, however, obtain a sort of "parity check" from writing permutations (nonuniquely) as products of 2-cycles.

Definition. A 2-cycle is called a *transposition*.

Intuitively, every permutation of $\{1, 2, \ldots, n\}$ can be realized by a succession of transpositions or simple interchanges of pairs of elements (try this on a small deck of cards sometime!). We illustrate how this may be done: first observe that

$$(a_1\, a_2 \ldots a_m) = (a_1\, a_m)(a_1\, a_{m-1})(a_1\, a_{m-2}) \ldots (a_1\, a_2)$$

for any m-cycle. Now any permutation in S_n may be written as a product of cycles (for instance, its cycle decomposition). Writing each of these cycles in turn as a product of transpositions by the above procedure we see that

every element of S_n may be written as a product of transpositions

or, equivalently,

$$S_n = \langle T \rangle \quad \text{where} \quad T = \{(i\,j) \mid 1 \le i < j \le n\}.$$

For example, the permutation σ in Section 1.3 may be written

$$\begin{aligned}\sigma &= (1\,12\,8\,10\,4)(2\,13)(5\,11\,7)(6\,9)\\ &= (1\,4)(1\,10)(1\,8)(1\,12)(2\,13)(5\,7)(5\,11)(6\,9).\end{aligned}$$

Again we emphasize that for any $\sigma \in S_n$ there may be many ways of writing σ as a product of transpositions. For fixed σ we now show that the parity (i.e. an odd or even number of terms) is the same for any product of transpositions equaling σ.

Let $x_1, \ldots, x_n$ be independent variables and let Δ be the polynomial:

$$\Delta = \prod_{1 \le i < j \le n} (x_i - x_j),$$

i.e. the product of all the terms $x_i - x_j$ for $i < j$. For example, when $n = 4$,

$$\Delta = (x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_2 - x_3)(x_2 - x_4)(x_3 - x_4).$$

For each $\sigma \in S_n$ let σ act on Δ by permuting the variables in the same way it permutes their indices:

$$\sigma(\Delta) = \prod_{1 \le i < j \le n} (x_{\sigma(i)} - x_{\sigma(j)}).$$

For example, if $n = 4$ and $\sigma = (1\ 2\ 3\ 4)$ then

$$\sigma(\Delta) = (x_2 - x_3)(x_2 - x_4)(x_2 - x_1)(x_3 - x_4)(x_3 - x_1)(x_4 - x_1)$$

(we have written the factors in the same order as above and applied σ to each factor to get $\sigma(\Delta)$). Note (in general) that Δ contains one factor $x_i - x_j$ for all $i < j$, and since σ is a bijection of the indices, $\sigma(\Delta)$ must contain either $x_i - x_j$ or $x_j - x_i$, but not both (and certainly no $x_i - x_i$ terms), for all $i < j$. If $\sigma(\Delta)$ has a factor $x_j - x_i$ where $j > i$, write this term as $-(x_i - x_j)$. Collecting all the changes in sign together we see that Δ and $\sigma(\Delta)$ have the same factors up to a product of -1's, i.e.

$$\sigma(\Delta) = \pm\Delta, \qquad \text{for all } \sigma \in S_n.$$

For each $\sigma \in S_n$ let

$$\epsilon(\sigma) = \begin{cases} +1, & \text{if } \sigma(\Delta) = \Delta \\ -1, & \text{if } \sigma(\Delta) = -\Delta. \end{cases}$$

In the example above with $n = 4$ and $\sigma = (1\ 2\ 3\ 4)$, there are exactly 3 factors of the form $x_j - x_i$ where $j > i$ in $\sigma(\Delta)$, each of which contributes a factor of -1. Hence

$$(1\ 2\ 3\ 4)(\Delta) = (-1)^3(\Delta) = -\Delta,$$

so

$$\epsilon((1\ 2\ 3\ 4)) = -1.$$

Definition

(1) $\epsilon(\sigma)$ is called the *sign* of σ.

(2) σ is called an *even permutation* if $\epsilon(\sigma) = 1$ and an *odd permutation* if $\epsilon(\sigma) = -1$.

We next prove:

Proposition 23. The map $\epsilon : S_n \to \{\pm 1\}$ is a homomorphism (where $\{\pm 1\}$ is a multiplicative version of the cyclic group of order 2).

Proof: By definition,

$$(\tau\sigma)(\Delta) = \prod_{1 \le i < j \le n} (x_{\tau\sigma(i)} - x_{\tau\sigma(j)}).$$

Suppose that $\sigma(\Delta)$ has exactly k factors of the form $x_j - x_i$ with $j > i$, that is $\epsilon(\sigma) = (-1)^k$. When calculating $(\tau\sigma)(\Delta)$, after first applying σ to the indices we see that $(\tau\sigma)(\Delta)$ has exactly k factors of the form $x_{\tau(j)} - x_{\tau(i)}$ with $j > i$. Interchanging the order of the terms in these k factors introduces the sign change $(-1)^k = \epsilon(\sigma)$, and now all factors of $(\tau\sigma)(\Delta)$ are of the form $x_{\tau(p)} - x_{\tau(q)}$, with $p < q$. Thus

$$(\tau\sigma)(\Delta) = \epsilon(\sigma) \prod_{1 \le p < q \le n} (x_{\tau(p)} - x_{\tau(q)}).$$

Since by definition of ϵ

$$\prod_{1 \le p < q \le n} (x_{\tau(p)} - x_{\tau(q)}) = \epsilon(\tau)\Delta$$

we have $(\tau\sigma)(\Delta) = \epsilon(\sigma)\epsilon(\tau)\Delta$. Thus $\epsilon(\tau\sigma) = \epsilon(\sigma)\epsilon(\tau) = \epsilon(\tau)\epsilon(\sigma)$, as claimed.

To see the proof in action, let $n = 4$, $\sigma = (1\,2\,3\,4)$, $\tau = (4\,2\,3)$ so $\tau\sigma = (1\,3\,2\,4)$. By definition (using the explicit Δ in this case),

$$\begin{aligned}(\tau\sigma)(\Delta) &= (1\,3\,2\,4)(\Delta)\\ &= (x_3 - x_4)(x_3 - x_2)(x_3 - x_1)(x_4 - x_2)(x_4 - x_1)(x_2 - x_1)\\ &= (-1)^5\Delta\end{aligned}$$

where all factors except the first one are flipped to recover Δ . This shows $\epsilon(\tau\sigma) = -1$. On the other hand, since we already computed $\sigma(\Delta)$

$$\begin{aligned}(\tau\sigma)(\Delta) &= \tau(\sigma(\Delta))\\ &= (x_{\tau(2)} - x_{\tau(3)})(x_{\tau(2)} - x_{\tau(4)})(x_{\tau(2)} - x_{\tau(1)})(x_{\tau(3)} - x_{\tau(4)}) \times\\ &\quad \times (x_{\tau(3)} - x_{\tau(1)})(x_{\tau(4)} - x_{\tau(1)})\\ &= (-1)^3 \prod_{1 \le p < q \le 4} (x_{\tau(p)} - x_{\tau(q)}) = (-1)^3\tau(\Delta)\end{aligned}$$

where here the third, fifth, and sixth factors need to have their terms interchanged in order to put all factors in the form $x_{\tau(p)} - x_{\tau(q)}$ with $p < q$. We already calculated that $\epsilon(\sigma) = (-1)^3 = -1$ and, by the same method, it is easy to see that $\epsilon(\tau) = (-1)^2 = 1$ so $\epsilon(\tau\sigma) = -1 = \epsilon(\tau)\epsilon(\sigma)$.

The next step is to compute $\epsilon((i\ j))$, for any transposition $(i\ j)$. Rather than compute this directly for arbitrary i and j we do it first for $i = 1$ and $j = 2$ and reduce

the general case to this. It is clear that applying $(1\ 2)$ to Δ (regardless of what n is) will flip exactly one factor, namely $x_1 - x_2$; thus $\epsilon((1\ 2)) = -1$. Now for any transposition $(i\ j)$ let λ be the permutation which interchanges 1 and i, interchanges 2 and j, and leaves all other numbers fixed (if $i = 1$ or $j = 2$, λ fixes i or j, respectively). Then it is easy to see that $(i\ j) = \lambda(1\ 2)\lambda$ (compute what the right hand side does to any $k \in \{1, 2, \ldots, n\}$). Since ϵ is a homomorphism we obtain

$$\begin{aligned}\epsilon((i\ j)) &= \epsilon(\lambda(1\ 2)\lambda)\\ &= \epsilon(\lambda)\epsilon((1\ 2))\epsilon(\lambda)\\ &= (-1)\epsilon(\lambda)^2\\ &= -1.\end{aligned}$$

This proves:

Proposition 24. Transpositions are all odd permutations and ϵ is a surjective homomorphism.

Definition. The *alternating group of degree n*, denoted by A_n, is the kernel of the homomorphism ϵ (i.e. the set of even permutations).

Note that by the First Isomorphism Theorem $S_n/A_n \cong \epsilon(S_n) = \{\pm 1\}$, so that the order of A_n is easily determined: $|A_n| = \frac{1}{2}|S_n| = \frac{1}{2}(n!)$. Also, $S_n - A_n$ is the coset of A_n which is not the identity coset and this is the set of all odd permutations. The signs of permutations obey the usual $\mathbb{Z}/2\mathbb{Z}$ laws:

$$\begin{aligned}(even)(even) &= (odd)(odd) = even\\ (even)(odd) &= (odd)(even) = odd.\end{aligned}$$

Moreover, since ϵ is a homomorphism and every $\sigma \in S_n$ is a product of transpositions, say $\sigma = \tau_1\tau_2\ldots\tau_k$, then $\epsilon(\sigma) = \epsilon(\tau_1)\ldots\epsilon(\tau_k)$; since $\epsilon(\tau_i) = -1$, for $i = 1, 2, \ldots, k$, $\epsilon(\sigma) = (-1)^k$. Thus the class of $k \pmod 2$ is the same no matter how we write σ as a product of transpositions:

$$\epsilon(\sigma) = \begin{cases} +1, & \text{if } \sigma \text{ is a product of an even number of transpositions}\\ -1, & \text{if } \sigma \text{ is a product of an odd number of transpositions.}\end{cases}$$

Finally we give a quick way of computing $\epsilon(\sigma)$ from the cycle decomposition of σ. Recall that an m-cycle may be written as a product of $m - 1$ transpositions. Thus

an m-cycle is an odd permutation if and only if m is even.

For any permutation σ let $\alpha_1\alpha_2\ldots\alpha_k$ be its cycle decomposition. Thus $\epsilon(\sigma) = \epsilon(\alpha_1)\ldots\epsilon(\alpha_k)$ and $\epsilon(\alpha_i) = -1$ if and only if the length of α_i is even. For $\epsilon(\sigma)$ to be -1 the product of the $\epsilon(\alpha_i)$'s must contain an odd number of factors of (-1). In summary:

Proposition 25. The permutation σ is odd if and only if the number of cycles of even length in its cycle decomposition is odd.

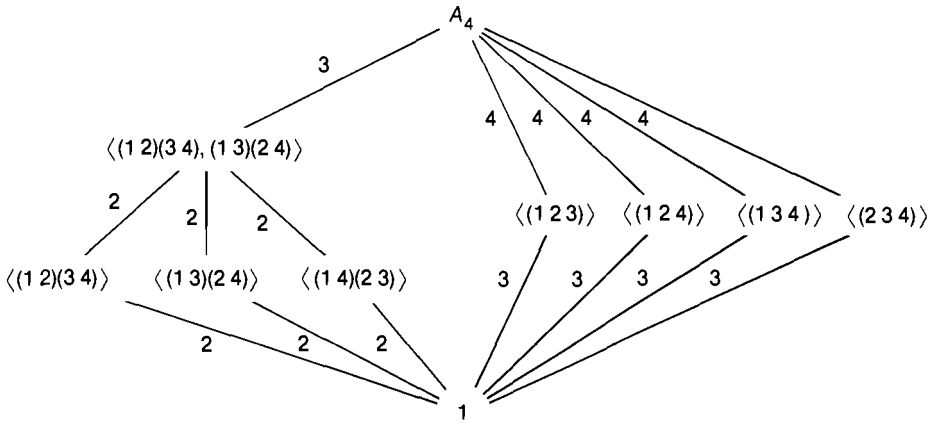


Fig. 8

For example, $\sigma = (1\ 12\ 8\ 10\ 4)(2\ 13)(5\ 11\ 7)(6\ 9)$ has exactly 2 cycles of even length, so $\epsilon(\sigma) = 1$. On the other hand, $\tau = (1\ 2\ 3\ 4\ 5\ 6)(7\ 8\ 9)(10\ 11)(12\ 13\ 14\ 15)(16\ 17\ 18)$ has 3 cycles of even length, hence $\epsilon(\tau) = -1$.

Be careful not to confuse the terms "odd" and "even" for a permutation σ with the parity of the order of σ. In fact, if σ is of odd order, all cycles in the cycle decomposition of σ have odd length so σ has an even (in this case 0) number of cycles of even length and hence is an even permutation. If $|\sigma|$ is even, σ may be either an even or an odd permutation; e.g. $(1\ 2)$ is odd, $(1\ 2)(3\ 4)$ is even but both have order 2.

As we mentioned in the preceding section, the alternating groups A_n will be important in the study of solvability of polynomials. In the next chapter we shall prove:

$$A_n \textit{ is a non-abelian simple group for all } n \geq 5.$$

For small values of n, A_n is already familiar to us: A_1 and A_2 are both the trivial group and $|A_3| = 3$ (so $A_3 = \langle(1\ 2\ 3)\rangle \cong Z_3$). The group A_4 has order 12. Exercise 7 shows A_4 is isomorphic to the group of symmetries of a regular tetrahedron. The lattice of subgroups of A_4 appears in Fig. 8 (Exercise 8 asserts that this is its complete lattice of subgroups). One of the nicer aspects of this lattice is that (unlike "virtually all groups") it is a planar graph (there are no crossing lines except at the vertices; see the lattice of D_{16} for a nonplanar lattice).

EXERCISES

1. In Exercises 1 and 2 of Section 1.3 you were asked to find the cycle decomposition of some permutations. Write each of these permutations as a product of transpositions. Determine which of these is an even permutation and which is an odd permutation.
2. Prove that σ^2 is an even permutation for every permutation σ.
3. Show that $S_n = \langle(1\ 2),\ (1\ 2\ 3 \ldots n)\rangle$ for all $n \geq 2$.
4. Show that if p is prime, $S_p = \langle\sigma, \tau\rangle$ where σ is any transposition and τ is any p-cycle.
5. Show that $\langle(1\ 3), (1\ 2\ 3\ 4)\rangle$ is a proper subgroup of S_4. What is the isomorphism type of this subgroup?

6. Prove that S_n is generated by $\{(i\ \ i+1) \mid 1 \leq i \leq n-1\}$.
7. Prove that the group of rigid motions of a tetrahedron is isomorphic to A_4. [Recall Exercise 20 in Section 1.7.]
8. Prove that the lattice of subgroups of A_4 given in the text is correct. [By the preceding exercise and the comments following Lagrange's Theorem, A_4 has no subgroup of order 6.]
9. Prove that the (unique) subgroup of order 4 in A_4 is normal and is isomorphic to V_4.
10. Find a composition series for A_4. Deduce that A_4 is solvable.
11. Prove that S_4 has no subgroup isomorphic to Q_8.
12. Prove that A_n contains a subgroup isomorphic to S_{n-2} for each $n \geq 3$.
13. Prove that every element of order 2 in A_n is the square of an element of order 4 in S_n. [An element of order 2 in A_n is a product of $2k$ commuting transpositions.]
14. Prove that the subgroup of A_4 generated by any element of order 2 and any element of order 3 is all of A_4.
15. Prove that if x and y are distinct 3-cycles in S_4 with $x \neq y^{-1}$, then the subgroup of S_4 generated by x and y is A_4.
16. Let x and y be distinct 3-cycles in S_5 with $x \neq y^{-1}$.
    (a) Prove that if x and y fix a common element of $\{1, \ldots, 5\}$, then $\langle x, y\rangle \cong A_4$.
    (b) Prove that if x and y do not fix a common element of $\{1, \ldots, 5\}$, then $\langle x, y\rangle = A_5$.
17. Prove that if x and y are 3-cycles in S_n, then $\langle x, y\rangle$ is isomorphic to either Z_3, A_4, A_5 or $Z_3 \times Z_3$.

4

Group Actions

In this chapter we consider some of the consequences of a group acting on a set. It is an important and recurring idea in mathematics that when one object acts on another then much information can be obtained on both. As more structure is added to the set on which the group acts (for example, groups acting on groups or groups acting on vector spaces (considered in Chapter 15)), more information on the structure of the group becomes available. This study of group actions culminates here in the proof of Sylow's Theorem and the examples and classifications which accrue from it.

The concept of an action will recur as we study modules, vector spaces, canonical forms for matrices and Galois Theory, and is one of the fundamental unifying themes in the text.

4.1 GROUP ACTIONS AND PERMUTATION REPRESENTATIONS

In this section we give the basic theory of group actions and then apply this theory to subgroups of S_n acting on $\{1, 2, \ldots, n\}$ to prove that every element of S_n has a unique cycle decomposition. In Sections 2 and 3 we apply the general theory to two other specific group actions to derive some important results.

Let G be a group acting on a nonempty set A. Recall from Section 1.7 that for each $g \in G$ the map

$$\sigma_g : A \to A \qquad \text{defined by} \qquad \sigma_g : a \mapsto g \cdot a$$

is a permutation of A. We also saw in Section 1.7 that there is a homomorphism associated to an action of G on A

$$\varphi : G \to S_A \qquad \text{defined by} \qquad \varphi(g) = \sigma_g,$$

called the *permutation representation* associated to the given action. Recall some additional terminology associated to group actions introduced in Sections 1.7 and 2.2.

Definition

(1) The *kernel* of the action is the set of elements of G that act trivially on every element of A: $\{g \in G \mid g \cdot a = a \text{ for all } a \in A\}$.

(2) For each $a \in A$ the *stabilizer* of a in G is the set of elements of G that fix the element a: $\{g \in G \mid g \cdot a = a\}$ and is denoted by G_a.

(3) An action is *faithful* if its kernel is the identity.

Note that the kernel of an action is precisely the same as the kernel of the associated permutation representation; in particular, the kernel is a normal subgroup of G. Two group elements induce the same permutation on A if and only if they are in the same coset of the kernel (if and only if they are in the same fiber of the permutation representation φ). In particular an action of G on A may also be viewed as a faithful action of the quotient group $G/\ker\varphi$ on A. Recall from Section 2.2 that the stabilizer in G of an element a of A is a subgroup of G. If a is a fixed element of A, then the kernel of the action is contained in the stabilizer G_a since the kernel of the action is the set of elements of G that stabilize every point, namely $\cap_{a\in A} G_a$.

Examples

(1) Let n be a positive integer. The group $G = S_n$ acts on the set $A = \{1, 2, \ldots, n\}$ by $\sigma \cdot i = \sigma(i)$ for all $i \in \{1, \ldots, n\}$. The permutation representation associated to this action is the identity map $\varphi : S_n \to S_n$. This action is faithful and for each $i \in \{1, \ldots, n\}$ the stabilizer G_i (the subgroup of all permutations fixing i) is isomorphic to S_{n-1} (cf. Exercise 15, Section 3.2).

(2) Let $G = D_8$ act on the set A consisting of the four vertices of a square. Label these vertices 1,2,3,4 in a clockwise fashion as in Fig. 2 of Section 1.2. Let r be the rotation of the square clockwise by $\pi/2$ radians and let s be the reflection in the line which passes through vertices 1 and 3. Then the permutations of the vertices given by r and s are

$$\sigma_r = (1\ 2\ 3\ 4) \qquad \text{and} \qquad \sigma_s = (2\ 4).$$

Note that since the permutation representation is a homomorphism, the permutation of the four vertices corresponding to sr is $\sigma_{sr} = \sigma_s\sigma_r = (1\ 4)(2\ 3)$. The action of D_8 on the four vertices of a square is faithful since only the identity symmetry fixes all four vertices. The stabilizer of any vertex a is the subgroup of D_8 of order 2 generated by the reflection about the line passing through a and the center of the square (so, for example, the stabilizer of vertex 1 is $\langle s \rangle$).

(3) Label the four vertices of a square as in the preceding example and now let A be the set whose elements consist of unordered pairs of opposite vertices: $A = \{\ \{1,3\}\ ,\ \{2,4\}\ \}$. Then D_8 also acts on this set A since each symmetry of the square sends a pair of opposite vertices to a pair of opposite vertices. The rotation r interchanges the pairs $\{1,3\}$ and $\{2,4\}$; the reflection s fixes both unordered pairs of opposite vertices. Thus if we label the pairs $\{1,3\}$ and $\{2,4\}$ as **1** and **2**, respectively, then the permutations of A given by r and s are

$$\sigma_r = (\mathbf{1}\ \mathbf{2}) \qquad \text{and} \qquad \sigma_s = \text{the identity permutation.}$$

This action of D_8 is not faithful: its kernel is $\langle s, r^2 \rangle$. Moreover, for each $a \in A$ the stabilizer in D_8 of a is the same as the kernel of the action.

(4) Label the four vertices of a square as in Example 2 and now let A be the following set of unordered pairs of vertices: $\{\ \{1,2\}\ ,\ \{3,4\}\ \}$. The group D_8 does *not* act on this set A because $\{1,2\} \in A$ but $r \cdot \{1,2\} = \{2,3\} \notin A$.

The relation between actions and homomorphisms into symmetric groups may be reversed. Namely, given any nonempty set A and any homomorphism φ of the group G into S_A we obtain an action of G on A by defining

$$g \cdot a = \varphi(g)(a)$$

for all $g \in G$ and all $a \in A$. The kernel of this action is the same as $\ker \varphi$. The permutation representation associated to this action is precisely the given homomorphism φ. This proves:

Proposition 1. For any group G and any nonempty set A there is a bijection between the actions of G on A and the homomorphisms of G into S_A.

In view of Proposition 1 the definition of a permutation representation may be rephrased.

Definition. If G is a group, a *permutation representation* of G is any homomorphism of G into the symmetric group S_A for some nonempty set A. We shall say a given action of G on A *affords* or *induces* the associated permutation representation of G.

We can think of a permutation representation as an analog of the matrix representation of a linear transformation. In the case where A is a finite set of n elements we have $S_A \cong S_n$ (cf. Section 1.6), so by fixing a labelling of the elements of A we may consider our permutations as elements of the group S_n (which is exactly what we did in Examples 2 and 3 above), in the same way that fixing a basis for a vector space allows us to view a linear transformation as a matrix.

We now prove a combinatorial result about group actions which will have important consequences when we apply it to specific actions in subsequent sections.

Proposition 2. Let G be a group acting on the nonempty set A. The relation on A defined by

$$a \sim b \quad \text{if and only if} \quad a = g \cdot b \text{ for some } g \in G$$

is an equivalence relation. For each $a \in A$, the number of elements in the equivalence class containing a is $|G : G_a|$, the index of the stabilizer of a.

Proof: We first prove $\sim$ is an equivalence relation. By axiom 2 of an action, $a = 1 \cdot a$ for all $a \in A$, i.e. $a \sim a$ and the relation is reflexive. If $a \sim b$, then $a = g \cdot b$ for some $b \in G$ so that

$$g^{-1} \cdot a = g^{-1} \cdot (g \cdot b) = (g^{-1}g) \cdot b = 1 \cdot b = b$$

that is, $b \sim a$ and the relation is symmetric. Finally, if $a \sim b$ and $b \sim c$, then $a = g \cdot b$ and $b = h \cdot c$, for some $g, h \in G$ so

$$a = g \cdot b = g \cdot (h \cdot c) = (gh) \cdot c$$

hence $a \sim c$, and the relation is transitive.

To prove the last statement of the proposition we exhibit a bijection between the left cosets of G_a in G and the elements of the equivalence class of a. Let $\mathcal{C}_a$ be the class of a, so

$$\mathcal{C}_a = \{g \cdot a \mid g \in G\}.$$

Suppose $b = g \cdot a \in \mathcal{C}_a$. Then gG_a is a left coset of G_a in G. The map

$$b = g \cdot a \mapsto gG_a$$

is a map from $\mathcal{C}_a$ to the set of left cosets of G_a in G. This map is surjective since for any $g \in G$ the element $g \cdot a$ is an element of $\mathcal{C}_a$. Since $g \cdot a = h \cdot a$ if and only if $h^{-1}g \in G_a$ if and only if $gG_a = hG_a$, the map is also injective, hence is a bijection. This completes the proof.

By Proposition 2 a group G acting on the set A partitions A into disjoint equivalence classes under the action of G. These classes are given a name:

Definition. Let G be a group acting on the nonempty set A.

(1) The equivalence class $\{g \cdot a \mid g \in G\}$ is called the *orbit* of G containing a.

(2) The action of G on A is called *transitive* if there is only one orbit, i.e. given any two elements $a, b \in A$ there is some $g \in G$ such that $a = g \cdot b$.

Examples

Let G be a group acting on the set A.

(1) If G acts trivially on A then $G_a = G$ for all $a \in A$ and the orbits are the elements of A. This action is transitive if and only if $|A| = 1$.

(2) The symmetric group $G = S_n$ acts transitively in its usual action as permutations on $A = \{1, 2, \ldots, n\}$. Note that the stabilizer in G of any point i has index $n = |A|$ in S_n.

(3) When the group G acts on the set A, any subgroup of G also acts on A. If G is transitive on A a subgroup of G need not be transitive on A. For example, if $G = \langle (1\ 2), (3\ 4) \rangle \leq S_4$ then the orbits of G on $\{1, 2, 3, 4\}$ are $\{1, 2\}$ and $\{3, 4\}$ and there is no element of G that sends 2 to 3. The discussion below on cycle decompositions shows that when $\langle \sigma \rangle$ is any cyclic subgroup of S_n then the orbits of $\langle \sigma \rangle$ consist of the sets of numbers that appear in the individual cycles in the cycle decomposition of σ (for example, the orbits of $\langle (1\ 2)(3\ 4\ 5) \rangle$ are $\{1, 2\}$ and $\{3, 4, 5\}$).

(4) The group D_8 acts transitively on the four vertices of the square and the stabilizer of any vertex is the subgroup of order 2 (and index 4) generated by the reflection about the line of symmetry passing through that point.

(5) The group D_8 also acts transitively on the set of two pairs of opposite vertices. In this action the stabilizer of any point is $\langle s, r^2 \rangle$ (which is of index 2).

Cycle Decompositions

We now prove that every element of the symmetric group S_n has the unique cycle decomposition described in Section 1.3. Let $A = \{1, 2, ..., n\}$, let σ be an element of S_n and let $G = \langle \sigma \rangle$. Then $\langle \sigma \rangle$ acts on A and so, by Proposition 2, it partitions $\{1, 2, \ldots, n\}$ into a unique set of (disjoint) orbits. Let $\mathcal{O}$ be one of these orbits and let $x \in \mathcal{O}$. By (the proof of) Proposition 2 applied to $A = \mathcal{O}$ we see that there is a bijection between the left cosets of G_x in G and the elements of $\mathcal{O}$, given explicitly by

$$\sigma^i x \mapsto \sigma^i G_x.$$

Since G is a cyclic group, $G_x \trianglelefteq G$ and G/G_x is cyclic of order d, where d is the smallest positive integer for which $\sigma^d \in G_x$ (cf. Example 2 following Proposition 7 in Section 3.1). Also, $d = |G : G_x| = |\mathcal{O}|$. Thus the distinct cosets of G_x in G are

$$1G_x,\ \sigma G_x,\ \sigma^2 G_x,\ \ldots,\ \sigma^{d-1} G_x.$$

This shows that the distinct elements of $\mathcal{O}$ are

$$x,\ \sigma(x),\ \sigma^2(x),\ \ldots,\ \sigma^{d-1}(x).$$

Ordering the elements of $\mathcal{O}$ in this manner shows that σ cycles the elements of $\mathcal{O}$, that is, on an orbit of size d, σ acts as a d-cycle. This proves the existence of a cycle decomposition for each $\sigma \in S_n$.

The orbits of $\langle\sigma\rangle$ are uniquely determined by σ. The only latitude is in which order the orbits are listed. Within each orbit, $\mathcal{O}$, we may begin with any element as a representative. Choosing $\sigma^i(x)$ instead of x as the initial representative simply produces the elements of $\mathcal{O}$ in the order

$$\sigma^i(x),\ \sigma^{i+1}(x),\ \ldots,\ \sigma^{d-1}(x),\ x,\ \sigma(x),\ \ldots,\ \sigma^{i-1}(x),$$

which is a cyclic permutation (forward $i-1$ terms) of the original list. It follows that the cycle decomposition above is unique up to a rearrangement of the cycles and up to a cyclic permutation of the integers within each cycle.

Subgroups of symmetric groups are called *permutation groups*. For any subgroup G of S_n the orbits of G will refer to its orbits on $\{1, 2, \ldots, n\}$. The orbits of an element σ in S_n will mean the orbits of the group $\langle\sigma\rangle$ (namely the sets of integers comprising the cycles in its cycle decomposition).

The exercises below further illustrate how group theoretic information can be obtained from permutation representations.

EXERCISES

Let G be a group and let A be a nonempty set.

1. Let G act on the set A. Prove that if $a, b \in A$ and $b = g \cdot a$ for some $g \in G$, then $G_b = gG_ag^{-1}$ (G_a is the stabilizer of a). Deduce that if G acts transitively on A then the kernel of the action is $\cap_{g\in G}\, gG_ag^{-1}$.
2. Let G be a *permutation group* on the set A (i.e. $G \le S_A$), let $\sigma \in G$ and let $a \in A$. Prove that $\sigma G_a \sigma^{-1} = G_{\sigma(a)}$. Deduce that if G acts transitively on A then
$$\bigcap_{\sigma\in G} \sigma G_a \sigma^{-1} = 1.$$
3. Assume that G is an abelian, transitive subgroup of S_A. Show that $\sigma(a) \ne a$ for all $\sigma \in G - \{1\}$ and all $a \in A$. Deduce that $|G| = |A|$. [Use the preceding exercise.]
4. Let S_3 act on the set Ω of ordered pairs: $\{(i, j) \mid 1 \le i, j \le 3\}$ by $\sigma((i, j)) = (\sigma(i), \sigma(j))$. Find the orbits of S_3 on Ω. For each $\sigma \in S_3$ find the cycle decomposition of σ under this action (i.e. find its cycle decomposition when σ is considered as an element of S_9 – first fix a labelling of these nine ordered pairs). For each orbit $\mathcal{O}$ of S_3 acting on these nine points pick some $a \in \mathcal{O}$ and find the stabilizer of a in S_3.
5. For each of parts (a) and (b) repeat the preceding exercise but with S_3 acting on the specified set:
   (a) the set of 27 triples $\{(i, j, k) \mid 1 \le i, j, k \le 3\}$
   (b) the set $\mathcal{P}(\{1, 2, 3\}) - \{\emptyset\}$ of all 7 nonempty subsets of $\{1, 2, 3\}$.
6. As in Exercise 12 of Section 2.2 let R be the set of all polynomials with integer coefficients in the independent variables x_1, x_2, x_3, x_4 and let S_4 act on R by permuting the indices of

the four variables:

$$\sigma \cdot p(x_1, x_2, x_3, x_4) = p(x_{\sigma(1)}, x_{\sigma(2)}, x_{\sigma(3)}, x_{\sigma(4)})$$

for all $\sigma \in S_4$.

(a) Find the polynomials in the orbit of S_4 on R containing $x_1 + x_2$ (recall from Exercise 12 in Section 2.2 that the stabilizer of this polynomial has order 4).

(b) Find the polynomials in the orbit of S_4 on R containing $x_1x_2 + x_3x_4$ (recall from Exercise 12 in Section 2.2 that the stabilizer of this polynomial has order 8).

(c) Find the polynomials in the orbit of S_4 on R containing $(x_1 + x_2)(x_3 + x_4)$.

7. Let G be a transitive permutation group on the finite set A. A *block* is a nonempty subset B of A such that for all $\sigma \in G$ either $\sigma(B) = B$ or $\sigma(B) \cap B = \emptyset$ (here $\sigma(B) = \{\sigma(b) \mid b \in B\}$).

(a) Prove that if B is a block containing the element a of A, then the set G_B defined by $G_B = \{\sigma \in G \mid \sigma(B) = B\}$ is a subgroup of G containing G_a.

(b) Show that if B is a block and $\sigma_1(B), \sigma_2(B), \ldots, \sigma_n(B)$ are all the distinct images of B under the elements of G, then these form a partition of A.

(c) A (transitive) group G on a set A is said to be *primitive* if the only blocks in A are the trivial ones: the sets of size 1 and A itself. Show that S_4 is primitive on $A = \{1, 2, 3, 4\}$. Show that D_8 is not primitive as a permutation group on the four vertices of a square.

(d) Prove that the transitive group G is primitive on A if and only if for each $a \in A$, the only subgroups of G containing G_a are G_a and G (i.e. G_a is a *maximal* subgroup of G: cf. Exercise 16, Section 2.4). [Use part (a).]

8. A transitive permutation group G on a set A is called *doubly transitive* if for any (hence all) $a \in A$ the subgroup G_a is transitive on the set $A - \{a\}$.

(a) Prove that S_n is doubly transitive on $\{1, 2, \ldots, n\}$ for all $n \geq 2$.

(b) Prove that a doubly transitive group is primitive. Deduce that D_8 is not doubly transitive in its action on the 4 vertices of a square.

9. Assume G acts transitively on the finite set A and let H be a normal subgroup of G. Let $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_r$ be the distinct orbits of H on A.

(a) Prove that G permutes the sets $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_r$ in the sense that for each $g \in G$ and each $i \in \{1, \ldots, r\}$ there is a j such that $g\mathcal{O}_i = \mathcal{O}_j$, where $g\mathcal{O} = \{g \cdot a \mid a \in \mathcal{O}\}$ (i.e. in the notation of Exercise 7 the sets $\mathcal{O}_1, \ldots, \mathcal{O}_r$ are blocks). Prove that G is transitive on $\{\mathcal{O}_1, \ldots, \mathcal{O}_r\}$. Deduce that all orbits of H on A have the same cardinality.

(b) Prove that if $a \in \mathcal{O}_1$ then $|\mathcal{O}_1| = |H : H \cap G_a|$ and prove that $r = |G : HG_a|$. [Draw the sublattice describing the Second Isomorphism Theorem for the subgroups H and G_a of G. Note that $H \cap G_a = H_a$.]

10. Let H and K be subgroups of the group G. For each $x \in G$ define the HK *double coset* of x in G to be the set

$$HxK = \{hxk \mid h \in H,\ k \in K\}.$$

(a) Prove that HxK is the union of the left cosets $x_1K, \ldots, x_nK$ where $\{x_1K, \ldots, x_nK\}$ is the orbit containing xK of H acting by left multiplication on the set of left cosets of K.

(b) Prove that HxK is a union of right cosets of H.

(c) Show that HxK and HyK are either the same set or are disjoint for all $x, y \in G$. Show that the set of HK double cosets partitions G.

(d) Prove that $|HxK| = |H| \cdot |H : H \cap xKx^{-1}|$.

(e) Prove that $|HxK| = |K| \cdot |K : K \cap x^{-1}Hx|$.

4.2 GROUPS ACTING ON THEMSELVES BY LEFT MULTIPLICATION — CAYLEY'S THEOREM

In this section G is any group and we first consider G *acting on itself (i.e. $A = G$) by left multiplication*:

$$g \cdot a = ga \qquad \text{for all } g \in G,\ a \in G$$

where ga denotes the product of the two group elements g and a in G (if G is written additively, the action will be written $g \cdot a = g + a$ and called left translation). We saw in Section 1.7 that this satisfies the two axioms of a group action.

When G is a finite group of order n it is convenient to label the elements of G with the integers $1, 2, \ldots, n$ in order to describe the permutation representation afforded by this action. In this way the elements of G are listed as $g_1, g_2, \ldots, g_n$ and for each $g \in G$ the permutation σ_g may be described as a permutation of the indices $1, 2, \ldots, n$ as follows:

$$\sigma_g(i) = j \qquad \text{if and only if} \qquad gg_i = g_j.$$

A different labelling of the group elements will give a different description of σ_g as a permutation of $\{1, 2, \ldots, n\}$ (cf. the exercises).

Example

Let $G = \{1, a, b, c\}$ be the Klein 4-group whose group table is written out in Section 2.5. Label the group elements $1, a, b, c$ with the integers 1,2,3,4 respectively. Under this labelling we compute the permutation σ_a induced by the action of left multiplication by the group element a:

$a \cdot 1 = a1 = a$ and so $\sigma_a(1) = 2$
$a \cdot a = aa = 1$ and so $\sigma_a(2) = 1$
$a \cdot b = ab = c$ and so $\sigma_a(3) = 4$ and
$a \cdot c = ac = b$ and so $\sigma_a(4) = 3$.

With this labelling of the elements of G we see that $\sigma_a = (1\ 2)(3\ 4)$. In the permutation representation associated to the action of the Klein 4-group on itself by left multiplication one similarly computes that

$$a \mapsto \sigma_a = (1\ 2)(3\ 4) \qquad b \mapsto \sigma_b = (1\ 3)(2\ 4) \qquad c \mapsto \sigma_c = (1\ 4)(2\ 3),$$

which explicitly gives the permutation representation $G \to S_4$ associated to this action under this labelling.

It is easy to see (and we shall prove this shortly in a more general setting) that the action of a group on itself by left multiplication is always transitive and faithful, and that the stabilizer of any point is the identity subgroup (these facts can be checked by inspection for the above example).

We now consider a generalization of the action of a group by left multiplication on the set of its elements. Let H be any subgroup of G and let A be the set of all left cosets of H in G. Define an action of G on A by

$$g \cdot aH = gaH \qquad \text{for all } g \in G,\ aH \in A$$

where gaH is the left coset with representative ga. One easily checks that this satisfies the two axioms for a group action, i.e. that G does act on the set of left cosets of H

by left multiplication. In the special case when H is the identity subgroup of G the coset aH is just $\{a\}$ and if we identify the element a with the set $\{a\}$, this action by left multiplication on left cosets of the identity subgroup is the same as the action of G on itself by left multiplication.

When H is of finite index m in G it is convenient to label the left cosets of H with the integers $1, 2, \ldots, m$ in order to describe the permutation representation afforded by this action. In this way the distinct left cosets of H in G are listed as $a_1H, a_2H, \ldots, a_mH$ and for each $g \in G$ the permutation σ_g may be described as a permutation of the indices $1, 2, \ldots, m$ as follows:

$$\sigma_g(i) = j \qquad \text{if and only if} \qquad ga_iH = a_jH.$$

A different labelling of the group elements will give a different description of σ_g as a permutation of $\{1, 2, \ldots, m\}$ (cf. the exercises).

Example

Let $G = D_8$ and let $H = \langle s \rangle$. Label the distinct left cosets $1H, rH, r^2H, r^3H$ with the integers 1,2,3,4 respectively. Under this labelling we compute the permutation σ_s induced by the action of left multiplication by the group element s on the left cosets of H:

$$\begin{aligned}
&s \cdot 1H = sH = 1H \text{ and so } \sigma_s(1) = 1\\
&s \cdot rH = srH = r^3H \text{ and so } \sigma_s(2) = 4\\
&s \cdot r^2H = sr^2H = r^2H \text{ and so } \sigma_s(3) = 3\\
&s \cdot r^3H = sr^3H = rH \text{ and so } \sigma_s(4) = 2.
\end{aligned}$$

With this labelling of the left cosets of H we obtain $\sigma_s = (2\ 4)$. In the permutation representation associated to the action of D_8 on the left cosets of $\langle s \rangle$ by left multiplication one similarly computes that $\sigma_r = (1\ 2\ 3\ 4)$. Note that the permutation representation is a homomorphism, so once its value has been determined on generators for D_8 its value on any other element can be determined (e.g. $\sigma_{sr^2} = \sigma_s\sigma_r^2$).

Theorem 3. Let G be a group, let H be a subgroup of G and let G act by left multiplication on the set A of left cosets of H in G. Let π_H be the associated permutation representation afforded by this action. Then

(1) G acts transitively on A

(2) the stabilizer in G of the point $1H \in A$ is the subgroup H

(3) the kernel of the action (i.e. the kernel of π_H) is $\cap_{x \in G}\, xHx^{-1}$, and $\ker \pi_H$ is the largest normal subgroup of G contained in H.

Proof: To see that G acts transitively on A let aH and bH be any two elements of A. Let $g = ba^{-1}$. Then $g \cdot aH = (ba^{-1})aH = bH$, and so the two arbitrary elements aH and bH of A lie in the same orbit. This proves (1).

The stabilizer of the point $1H$ is, by definition, $\{g \in G \mid g \cdot 1H = 1H\} = \{g \in G \mid gH = H\} = H$. This gives (2).

By definition of π_H we have

$$\begin{aligned}\ker \pi_H &= \{g \in G \mid gxH = xH \text{ for all } x \in G\}\\ &= \{g \in G \mid (x^{-1}gx)H = H \text{ for all } x \in G\}\\ &= \{g \in G \mid x^{-1}gx \in H \text{ for all } x \in G\}\\ &= \{g \in G \mid g \in xHx^{-1} \text{ for all } x \in G\}\\ &= \bigcap_{x \in G} xHx^{-1}.\end{aligned}$$

The second assertion of part (3) comes from observing first that $\ker \pi_H \trianglelefteq G$ and $\ker \pi_H \leq H$. If now N is any normal subgroup of G contained in H then $N = xNx^{-1} \leq xHx^{-1}$ for all $x \in G$ so that

$$N \leq \bigcap_{x \in G} xHx^{-1} = \ker \pi_H.$$

This shows that $\ker \pi_H$ is the largest normal subgroup of G contained in H.

Corollary 4. (Cayley's Theorem) Every group is isomorphic to a subgroup of some symmetric group. If G is a group of order n, then G is isomorphic to a subgroup of S_n.

Proof: Let $H = 1$ and apply the preceding theorem to obtain a homomorphism of G into S_G (here we are identifying the cosets of the identity subgroup with the elements of G). Since the kernel of this homomorphism is contained in $H = 1$, G is isomorphic to its image in S_G.

Note that G is isomorphic to a *subgroup* of a symmetric group, not to the full symmetric group itself. For example, we exhibited an isomorphism of the Klein 4-group with the subgroup $\langle (1\ 2)(3\ 4), (1\ 3)(2\ 4)\rangle$ of S_4. Recall that subgroups of symmetric groups are called *permutation groups* so Cayley's Theorem states that every group is isomorphic to a permutation group. The permutation representation afforded by left multiplication on the elements of G (cosets of $H = 1$) is called the *left regular representation* of G. One might think that we could study all groups more effectively by simply studying subgroups of symmetric groups (and all finite groups by studying subgroups of S_n, for all n). This approach alone is neither computationally nor theoretically practical, since to study groups of order n we would have to work in the much larger group S_n (cf. Exercise 7, for example).

Historically, finite groups were first studied not in an axiomatic setting as we have developed but as subgroups of S_n. Thus Cayley's Theorem proves that the historical notion of a group and the modern (axiomatic) one are equivalent. One advantage of the modern approach is that we are not, in our study of a given group, restricted to considering that group as a subgroup of some *particular* symmetric group (so in some sense our groups are "coordinate free").

The next result generalizes our result on the normality of subgroups of index 2.

Corollary 5. If G is a finite group of order n and p is the smallest prime dividing $|G|$, then any subgroup of index p is normal.

Remark: In general, a group of order n need not have a subgroup of index p (for example, A_4 has no subgroup of index 2).

Proof: Suppose $H \leq G$ and $|G : H| = p$. Let π_H be the permutation representation afforded by multiplication on the set of left cosets of H in G, let $K = \ker \pi_H$ and let $|H : K| = k$. Then $|G : K| = |G : H||H : K| = pk$. Since H has p left cosets, G/K is isomorphic to a subgroup of S_p (namely, the image of G under π_H) by the First Isomorphism Theorem. By Lagrange's Theorem, $pk = |G/K|$ divides $p!$. Thus $k \mid \dfrac{p!}{p} = (p-1)!$. But all prime divisors of $(p-1)!$ are less than p and by the minimality of p, every prime divisor of k is greater than or equal to p. This forces $k = 1$, so $H = K \trianglelefteq G$, completing the proof.

EXERCISES

Let G be a group and let H be a subgroup of G.

1. Let $G = \{1, a, b, c\}$ be the Klein 4-group whose group table is written out in Section 2.5.
   (a) Label $1, a, b, c$ with the integers 1,2,4,3 respectively and prove that under the left regular representation of G into S_4 the nonidentity elements are mapped as follows:
   $$a \mapsto (1\ 2)(3\ 4) \qquad b \mapsto (1\ 4)(2\ 3) \qquad c \mapsto (1\ 3)(2\ 4).$$
   (b) Relabel $1, a, b, c$ as 1,4,2,3 respectively and compute the image of each element of G under the left regular representation of G into S_4. Show that the image of G in S_4 under this labelling is the same *subgroup* as the image of G in part (a) (even though the nonidentity elements individually map to different permutations under the two different labellings).
2. List the elements of S_3 as 1, (1 2), (2 3), (1 3), (1 2 3), (1 3 2) and label these with the integers 1,2,3,4,5,6 respectively. Exhibit the image of each element of S_3 under the left regular representation of S_3 into S_6.
3. Let r and s be the usual generators for the dihedral group of order 8.
   (a) List the elements of D_8 as $1, r, r^2, r^3, s, sr, sr^2, sr^3$ and label these with the integers $1, 2, \ldots, 8$ respectively. Exhibit the image of each element of D_8 under the left regular representation of D_8 into S_8.
   (b) Relabel this same list of elements of D_8 with the integers $1, 3, 5, 7, 2, 4, 6, 8$ respectively and recompute the image of each element of D_8 under the left regular representation with respect to this new labelling. Show that the two subgroups of S_8 obtained in parts (a) and (b) are different.
4. Use the left regular representation of Q_8 to produce two elements of S_8 which generate a subgroup of S_8 isomorphic to the quaternion group Q_8.
5. Let r and s be the usual generators for the dihedral group of order 8 and let $H = \langle s \rangle$. List the left cosets of H in D_8 as $1H, rH, r^2H$ and r^3H.
   (a) Label these cosets with the integers 1,2,3,4 respectively. Exhibit the image of each element of D_8 under the representation π_H of D_8 into S_4 obtained from the action of D_8 by left multiplication on the set of 4 left cosets of H in D_8. Deduce that this rep-

resentation is faithful (i.e. the elements of S_4 obtained form a subgroup isomorphic to D_8).

(b) Repeat part (a) with the list of cosets relabelled by the integers 1,3,2,4 respectively. Show that the permutations obtained from this labelling form a subgroup of S_4 that is different from the subgroup obtained in part (a).

(c) Let $K = \langle sr \rangle$, list the cosets of K in D_8 as $1K$, rK, r^2K and r^3K, and label these with the integers 1,2,3,4. Prove that, with respect to this labelling, the image of D_8 under the representation π_K obtained from left multiplication on the cosets of K is the same *subgroup* of S_4 as in part (a) (even though the subgroups H and K are different and some of the elements of D_8 map to different permutations under the two homomorphisms).

6. Let r and s be the usual generators for the dihedral group of order 8 and let $N = \langle r^2 \rangle$. List the left cosets of N in D_8 as $1N$, rH, sN and srN. Label these cosets with the integers 1,2,3,4 respectively. Exhibit the image of each element of D_8 under the representation π_N of D_8 into S_4 obtained from the action of D_8 by left multiplication on the set of 4 left cosets of N in D_8. Deduce that this representation is not faithful and prove that $\pi_N(D_8)$ is isomorphic to the Klein 4-group.

7. Let Q_8 be the quaternion group of order 8.

(a) Prove that Q_8 is isomorphic to a subgroup of S_8.

(b) Prove that Q_8 is not isomorphic to a subgroup of S_n for any $n \leq 7$. [If Q_8 acts on any set A of order ≤ 7 show that the stabilizer of any point $a \in A$ must contain the subgroup $\langle -1 \rangle$.]

8. Prove that if H has finite index n then there is a normal subgroup K of G with $K \leq H$ and $|G : K| \leq n!$.

9. Prove that if p is a prime and G is a group of order p^α for some $\alpha \in \mathbb{Z}^+$, then every subgroup of index p is normal in G. Deduce that every group of order p^2 has a normal subgroup of order p.

10. Prove that every non-abelian group of order 6 has a nonnormal subgroup of order 2. Use this to classify groups of order 6. [Produce an injective homomorphism into S_3.]

11. Let G be a finite group and let $\pi : G \to S_G$ be the left regular representation. Prove that if x is an element of G of order n and $|G| = mn$, then $\pi(x)$ is a product of m n-cycles. Deduce that $\pi(x)$ is an odd permutation if and only if $|x|$ is even and $\dfrac{|G|}{|x|}$ is odd.

12. Let G and π be as in the preceding exercise. Prove that if $\pi(G)$ contains an odd permutation then G has a subgroup of index 2. [Use Exercise 3 in Section 3.3.]

13. Prove that if $|G| = 2k$ where k is odd then G has a subgroup of index 2. [Use Cauchy's Theorem to produce an element of order 2 and then use the preceding two exercises.]

14. Let G be a finite group of composite order n with the property that G has a subgroup of order k for each positive integer k dividing n. Prove that G is not simple.

4.3 GROUPS ACTING ON THEMSELVES BY CONJUGATION — THE CLASS EQUATION

In this section G is any group and we first consider G *acting on itself* (i.e. $A = G$) *by conjugation*:

$$g \cdot a = gag^{-1} \qquad \text{for all } g \in G,\ a \in G$$

where gag^{-1} is computed in the group G as usual. This definition satisfies the two axioms for a group action because:

$$\begin{aligned} g_1 \cdot (g_2 \cdot a) = g_1 \cdot (g_2 a g_2^{-1}) &= g_1(g_2 a g_2^{-1})g_1^{-1} \\ &= (g_1 g_2)a(g_1 g_2)^{-1} = (g_1 g_2) \cdot a \end{aligned}$$

and

$$1 \cdot a = 1a1^{-1} = a$$

for all $g_1, g_2 \in G$ and all $a \in G$.

Definition. Two elements a and b of G are said to be *conjugate in* G if there is some $g \in G$ such that $b = gag^{-1}$ (i.e. if and only if they are in the same orbit of G acting on itself by conjugation). The orbits of G acting on itself by conjugation are called the *conjugacy classes of* G.

Examples

(1) If G is an abelian group then the action of G on itself by conjugation is the trivial action: $g \cdot a = a$, for all $g, a \in G$, and for each $a \in G$ the conjugacy class of a is $\{a\}$.

(2) If $|G| > 1$ then, unlike the action by left multiplication, G does *not* act transitively on itself by conjugation because $\{1\}$ is always a conjugacy class (i.e. an orbit for this action). More generally, the one element subset $\{a\}$ is a conjugacy class if and only if $gag^{-1} = a$ for all $g \in G$ if and only if a is in the center of G.

(3) In S_3 one can compute directly that the conjugacy classes are: $\{1\}, \{(1\ 2), (1\ 3), (2\ 3)\}$ and $\{(1\ 2\ 3), (1\ 3\ 2)\}$. We shall shortly develop techniques for computing conjugacy classes more easily, particularly in symmetric groups.

As in the case of a group acting on itself by left multiplication, the action by conjugation can be generalized. If S is any subset of G, define

$$gSg^{-1} = \{gsg^{-1} \mid s \in S\}.$$

A group G acts on the set $\mathcal{P}(G)$ of all subsets of itself by defining $g \cdot S = gSg^{-1}$ for any $g \in G$ and $S \in \mathcal{P}(G)$. As above, this defines a group action of G on $\mathcal{P}(G)$. Note that if S is the one element set $\{s\}$ then $g \cdot S$ is the one element set $\{gsg^{-1}\}$ and so this action of G on all subsets of G may be considered as an extension of the action of G on itself by conjugation.

Definition. Two subsets S and T of G are said to be *conjugate in* G if there is some $g \in G$ such that $T = gSg^{-1}$ (i.e. if and only if they are in the same orbit of G acting on its subsets by conjugation).

We now apply Proposition 2 to the action of G by conjugation. Proposition 2 proves that if S is a subset of G, then the number of conjugates of S equals the index $|G : G_S|$ of the stabilizer G_S of S. For action by conjugation

$$G_S = \{g \in G \mid gSg^{-1} = S\} = N_G(S)$$

is the normalizer of S in G. We summarize this as

Proposition 6. The number of conjugates of a subset S in a group G is the index of the normalizer of S, $|G : N_G(S)|$. In particular, the number of conjugates of an element s of G is the index of the centralizer of s, $|G : C_G(s)|$.

Proof: The second assertion of the proposition follows from the observation that $N_G(\{s\}) = C_G(s)$.

The action of G on itself by conjugation partitions G into the conjugacy classes of G, whose orders can be computed by Proposition 6. Since the sum of the orders of these conjugacy classes is the order of G, we obtain the following important relation among these orders:

Theorem 7. (The Class Equation) Let G be a finite group and let $g_1, g_2, \ldots, g_r$ be representatives of the distinct conjugacy classes of G not contained in the center $Z(G)$ of G. Then

$$|G| = |Z(G)| + \sum_{i=1}^{r} |G : C_G(g_i)|.$$

Proof: As noted in Example 2 above the element $\{x\}$ is a conjugacy class of size 1 if and only if $x \in Z(G)$, since then $g^{-1}xg = x$ for all $g \in G$. Let $Z(G) = \{1, z_2, \ldots, z_m\}$, let $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_r$ be the conjugacy classes of G not contained in the center, and let g_i be a representative of $\mathcal{K}_i$ for each i. Then the full set of conjugacy classes of G is given by

$$\{1\},\ \{z_2\},\ \ldots,\ \{z_m\},\ \mathcal{K}_1,\ \mathcal{K}_2,\ \ldots,\ \mathcal{K}_r.$$

Since these partition G we have

$$\begin{aligned} |G| &= \sum_{i=1}^{m} 1 + \sum_{i=1}^{r} |\mathcal{K}_i| \\ &= |Z(G)| + \sum_{i=1}^{r} |G : C_G(g_i)|, \end{aligned}$$

where $|\mathcal{K}_i|$ is given by Proposition 6. This proves the class equation.

Note in particular that all the summands on the right hand side of the class equation are divisors of the group order since they are indices of subgroups of G. This restricts their possible values (cf. Exercise 6, for example).

Examples

(1) The class equation gives no information in an abelian group since conjugation is the trivial action and all conjugacy classes have size 1.

(2) In any group G we have $\langle g \rangle \leq C_G(g)$; this observation helps to minimize computations of conjugacy classes. For example, in the quaternion group Q_8 we see that $\langle i \rangle \leq C_{Q_8}(i) \leq Q_8$. Since $i \notin Z(Q_8)$ and $|Q_8 : \langle i \rangle| = 2$, we must have $C_{Q_8}(i) = \langle i \rangle$. Thus i has precisely 2 conjugates in Q_8, namely i and $-i = kik^{-1}$.

The other conjugacy classes in Q_8 are determined similarly and are:

$$\{1\}, \quad \{-1\}, \quad \{\pm i\}, \quad \{\pm j\}, \quad \{\pm k\}.$$

The first two classes form $Z(Q_8)$ and the class equation for this group is

$$|Q_8| = 2 + 2 + 2 + 2.$$

(3) In D_8 we may also use the fact that the three subgroups of index 2 are abelian to quickly see that if $x \notin Z(D_8)$, then $|C_{D_8}(x)| = 4$. The conjugacy classes of D_8 are:

$$\{1\}, \quad \{r^2\}, \quad \{r, r^3\}, \quad \{s, sr^2\}, \quad \{sr, sr^3\}.$$

The first two classes form $Z(D_8)$ and the class equation for this group is

$$|D_8| = 2 + 2 + 2 + 2.$$

Before discussing more examples of conjugacy we give two important consequences of the class equation. The first application of the class equation is to show that groups of prime power order have nontrivial centers, which is the starting point for the study of groups of prime power order (to which we return in Chapter 6).

Theorem 8. If p is a prime and P is a group of prime power order p^α for some $\alpha \geq 1$, then P has a nontrivial center: $Z(P) \neq 1$.

Proof: By the class equation

$$|P| = |Z(P)| + \sum_{i=1}^{r} |P : C_P(g_i)|$$

where $g_1, \ldots, g_r$ are representatives of the distinct non-central conjugacy classes. By definition, $C_P(g_i) \neq P$ for $i = 1, 2, \ldots, r$ so p divides $|P : C_P(g_i)|$. Since p also divides $|P|$ it follows that p divides $|Z(P)|$, hence the center must be nontrivial.

Corollary 9. If $|P| = p^2$ for some prime p, then P is abelian. More precisely, P is isomorphic to either Z_{p^2} or $Z_p \times Z_p$.

Proof: Since $Z(P) \neq 1$ by the theorem, it follows that $P/Z(P)$ is cyclic. By Exercise 36, Section 3.1, P is abelian. If P has an element of order p^2, then P is cyclic. Assume therefore that every nonidentity element of P has order p. Let x be any nonidentity element of P and let $y \in P - \langle x \rangle$. Since $|\langle x, y \rangle| > |\langle x \rangle| = p$, we must have that $P = \langle x, y \rangle$. Both x and y have order p so $\langle x \rangle \times \langle y \rangle = Z_p \times Z_p$. It now follows directly that the map $(x^a, y^b) \mapsto x^a y^b$ is an isomorphism from $\langle x \rangle \times \langle y \rangle$ onto P. This completes the proof.

Conjugacy in S_n

We next consider conjugation in symmetric groups. Readers familiar with linear algebra will recognize that in the matrix group $GL_n(F)$, conjugation is the same as "change of basis": $A \mapsto PAP^{-1}$. The situation in S_n is analogous:

Proposition 10. Let σ, τ be elements of the symmetric group S_n and suppose σ has cycle decomposition

$$(a_1\, a_2 \ldots a_{k_1})\,(b_1\, b_2 \ldots b_{k_2}) \ldots .$$

Then $\tau\sigma\tau^{-1}$ has cycle decomposition

$$(\,\tau(a_1)\,\tau(a_2) \ldots \tau(a_{k_1}))\,(\,\tau(b_1)\,\tau(b_2) \ldots \tau(b_{k_2}))\ \ldots$$

that is, $\tau\sigma\tau^{-1}$ is obtained from σ by replacing each entry i in the cycle decomposition for σ by the entry $\tau(i)$.

Proof: Observe that if $\sigma(i) = j$, then

$$\tau\sigma\tau^{-1}(\tau(i)) = \tau(j).$$

Thus, if the ordered pair i, j appears in the cycle decomposition of σ, then the ordered pair $\tau(i), \tau(j)$ appears in the cycle decomposition of $\tau\sigma\tau^{-1}$. This completes the proof.

Example

Let $\sigma = (1\,2)(3\,4\,5)(6\,7\,8\,9)$ and let $\tau = (1\,3\,5\,7)(2\,4\,6\,8)$. Then

$$\tau\sigma\tau^{-1} = (3\,4)(5\,6\,7)(8\,1\,2\,9)$$

Definition

(1) If $\sigma \in S_n$ is the product of disjoint cycles of lengths $n_1, n_2, \ldots, n_r$ with $n_1 \leq n_2 \leq \cdots \leq n_r$ (including its 1-cycles) then the integers $n_1, n_2, \ldots, n_r$ are called the *cycle type* of σ.

(2) If $n \in \mathbb{Z}^+$, a *partition* of n is any nondecreasing sequence of positive integers whose sum is n.

Note that by the results of the preceding section the cycle type of a permutation is unique. For example, the cycle type of an m-cycle in S_n is $1, 1, \ldots, 1, m$, where the m is preceded by $n - m$ ones.

Proposition 11. Two elements of S_n are conjugate in S_n if and only if they have the same cycle type. The number of conjugacy classes of S_n equals the number of partitions of n.

Proof: By Proposition 10, conjugate permutations have the same cycle type. Conversely, suppose the permutations σ_1 and σ_2 have the same cycle type. Order the cycles in nondecreasing length, including 1-cycles (if several cycles of σ_1 and σ_2 have the same length then there are several ways of doing this). Ignoring parentheses, each cycle decomposition is a list in which all the integers from 1 to n appear exactly once. Define τ to be the function which maps the ith integer in the list for σ_1 to the ith integer in the list for σ_2. Thus τ is a permutation and since the parentheses which delineate the cycle decompositions appear at the same positions in each list, Proposition 10 ensures that $\tau\sigma_1\tau^{-1} = \sigma_2$, so that σ_1 and σ_2 are conjugate.

Since there is a bijection between the conjugacy classes of S_n and the permissible cycle types and each cycle type for a permutation in S_n is a partition of n, the second assertion of the proposition follows, completing the proof.

Examples

(1) Let $\sigma_1 = (1)(3\,5)(8\,9)(2\,4\,7\,6)$ and let $\sigma_2 = (3)(4\,7)(8\,1)(5\,2\,6\,9)$. Then define τ by $\tau(1) = 3$, $\tau(3) = 4$, $\tau(5) = 7$, $\tau(8) = 8$, etc. Then

$$\tau = (1\,3\,4\,2\,5\,7\,6\,9)(8)$$

and $\tau\sigma_1\tau^{-1} = \sigma_2$.

(2) If in the previous example we had reordered σ_2 as $\sigma_2 = (3)(8\,1)(4\,7)(5\,2\,6\,9)$ by interchanging the two cycles of length 2, then the corresponding τ described above is defined by $\tau(1) = 3$, $\tau(3) = 8$, $\tau(5) = 1$, $\tau(8) = 4$ etc., which gives the permutation

$$\tau = (1\,3\,8\,4\,2\,5)(6\,9\,7)$$

again with $\tau\sigma_1\tau^{-1} = \sigma_2$, which shows that there are many elements conjugating σ_1 into σ_2.

(3) If $n = 5$, the partitions of 5 and corresponding representatives of the conjugacy classes (with 1-cycles not written) are as given in the following table:

Partition of 5	Representative of conjugacy class
1, 1, 1, 1, 1	1
1, 1, 1, 2	(1 2)
1, 1, 3	(1 2 3)
1,4	(1 2 3 4)
5	(1 2 3 4 5)
1, 2, 2	(1 2) (3 4)
2, 3	(1 2) (3 4 5)

Proposition 11 and Proposition 6 can be used to exhibit the centralizers of some elements in S_n. For example, if σ is an m-cycle in S_n, then the number of conjugates of σ (i.e. the number of m-cycles) is

$$\frac{n \cdot (n-1) \cdots (n-m+1)}{m}.$$

By Proposition 6 this is the index of the centralizer of σ: $\dfrac{|S_n|}{|C_{S_n}(\sigma)|}$. Since $|S_n| = n!$ we obtain

$$|C_{S_n}(\sigma)| = m \cdot (n-m)!.$$

The element σ certainly commutes with $1, \sigma, \sigma^2, \ldots, \sigma^{m-1}$. It also commutes with any permutation in S_n whose cycles are disjoint from σ and there are $(n-m)!$ permutations of this type (the full symmetric group on the numbers not appearing in σ). The product of elements of these two types already accounts for $m \cdot (n-m)!$ elements commuting with σ. By the order computation above, this is the full centralizer of σ in S_n. Explicitly,

if σ is an m-cycle in S_n, then $C_{S_n}(\sigma) = \{\sigma^i\tau \mid 0 \le i \le m-1,\ \tau \in S_{n-m}\}$

where S_{n-m} denotes the subgroup of S_n which fixes all integers appearing in the m-cycle σ (and is the identity subgroup if $m = n$ or $m = n - 1$).

For example, the centralizer of $\sigma = (1\,3\,5)$ in S_7 is the subgroup

$$\{(1\,3\,5)^i\tau \mid i = 0, 1 \text{ or } 2, \text{ and } \tau \text{ fixes } 1, 3 \text{ and } 5\}.$$

Note that $\tau \in S_A$ where $A = \{2, 4, 6, 7\}$, so there are 4! choices for τ and the centralizer has order $3 \cdot 4! = 72$.

We shall discuss centralizers of other elements of S_n in the next exercises and in Chapter 5.

We can use this discussion of the conjugacy classes in S_n to give a combinatorial proof of the simplicity of A_5. We first observe that normal subgroups of a group G are the union of conjugacy classes of G, i.e.

if $H \trianglelefteq G$, then for every conjugacy class $\mathcal{K}$ of G either $\mathcal{K} \subseteq H$ or $\mathcal{K} \cap H = \emptyset$.

This is because if $x \in \mathcal{K} \cap H$, then $gxg^{-1} \in gHg^{-1}$ for all $g \in G$. Since H is normal, $gHg^{-1} = H$, so that H contains all the conjugates of x, i.e. $\mathcal{K} \subseteq H$.

Theorem 12. A_5 is a simple group.

Proof: We first work out the conjugacy classes of A_5 and their orders. Proposition 11 does not apply directly since two elements of the same cycle type (which are conjugate in S_5) need *not* be conjugate in A_5. Exercises 19 to 22 analyze the relation of classes in S_n to classes in A_n in detail.

We have already seen that representatives of the cycle types of even permutations can be taken to be

$$1, \quad (1\,2\,3), \quad (1\,2\,3\,4\,5) \quad \text{and} \quad (1\,2)(3\,4).$$

The centralizers of 3-cycles and 5-cycles in S_5 were determined above, and checking which of these elements are contained in A_5 we see that

$$C_{A_5}((1\,2\,3)) = \langle(1\,2\,3)\rangle \quad \text{and} \quad C_{A_5}((1\,2\,3\,4\,5)) = \langle(1\,2\,3\,4\,5)\rangle.$$

These groups have orders 3 and 5 (index 20 and 12), respectively, so there are 20 distinct conjugates of (1 2 3) and 12 distinct conjugates of (1 2 3 4 5) in A_5. Since there are a total of twenty 3-cycles in S_5 (Exercise 16, Section 1.3) and all of these lie in A_5, we see that

all twenty 3-cycles are conjugate in A_5.

There are a total of twenty-four 5-cycles in A_5 but only 12 distinct conjugates of the 5-cycle (1 2 3 4 5). Thus some 5-cycle, σ, is *not* conjugate to (1 2 3 4 5) in A_5 (in fact, (1 3 5 2 4) is not conjugate in A_5 to (1 2 3 4 5) since the method of proof in Proposition 11 shows that any element of S_5 conjugating (1 2 3 4 5) into (1 3 5 2 4) must be an odd permutation). As above we see that σ also has 12 distinct conjugates in A_5, hence

the 5-cycles lie in two conjugacy classes in A_5, each of which has 12 elements.

Since the 3-cycles and 5-cycles account for all the nonidentity elements of odd order, the 15 remaining nonidentity elements of A_5 must have order 2 and therefore have

cycle type (2,2). It is easy to see that (1 2)(3 4) commutes with (1 3)(2 4) but does not commute with any element of odd order in A_5. It follows that $|C_{A_5}((12)(34))| = 4$. Thus (1 2)(3 4) has 15 distinct conjugates in A_5, hence

all 15 elements of order 2 in A_5 are conjugate to (1 2)(3 4).

In summary, the conjugacy classes of A_5 have orders 1, 15, 20, 12 and 12.

Now, suppose H were a normal subgoup of A_5. Then as we observed above, H would be the union of conjugacy classes of A_5. Then the order of H would be both a divisor of 60 (the order of A_5) and be the sum of some collection of the integers $\{1, 12, 12, 15, 20\}$ (the sizes of the conjugacy classes in A_5). A quick check shows the only possibilities are $|H| = 1$ or $|H| = 60$, so that A_5 has no proper, nontrivial normal subgroups.

Right Group Actions

As noted in Section 1.7, in the definition of an action the group elements appear to the left of the set elements and so our notion of an action might more precisely be termed a *left group action*. One can analogously define the notion of a *right group action* of the group G on the nonempty set A as a map from $A \times G$ to A, denoted by $a \cdot g$ for $a \in A$ and $g \in G$, that satisfies the axioms:

(1) $(a \cdot g_1) \cdot g_2 = a \cdot (g_1 g_2)$ for all $a \in A$, and $g_1, g_2 \in G$, and
(2) $a \cdot 1 = a$ for all $a \in A$.

In much of the literature on group theory, conjugation is written as a right group action using the following notation:

$$a^g = g^{-1}ag \qquad \text{for all } g, a \in G.$$

Similarly, for subsets S of G one defines $S^g = g^{-1}Sg$. In this notation the two axioms for a right action are verified as follows:

$$(a^{g_1})^{g_2} = g_2^{-1}(g_1^{-1}ag_1)g_2 = (g_1g_2)^{-1}a(g_1g_2) = a^{(g_1g_2)}$$

and

$$a^1 = 1^{-1}a1 = a$$

for all $g_1, g_2, a \in G$. Thus the two axioms for this right action of a group on itself take the form of the familiar "laws of exponentiation". (Note that the integer power a^n of a group element a is easily distinguished from the conjugate a^g of a by the nature of the exponent: $n \in \mathbb{Z}$ but $g \in G$.) Because conjugation is so ubiquitous in the theory of groups, this notation is a useful and efficient shorthand (as opposed to always writing gag^{-1} or $g \cdot a$ for action on the left by conjugation).

For arbitrary group actions it is an easy exercise to check that if we are given a left group action of G on A then the map $A \times G \to A$ defined by $a \cdot g = g^{-1} \cdot a$ is a right group action. Conversely, given a right group action of G on A we can form a left group action by $g \cdot a = a \cdot g^{-1}$. Call these pairs *corresponding group actions*. Put another way, for corresponding group actions, g acts on the left in the same way that

g^{-1} acts on the right. This is particularly transparent for the action of conjugation because the "left conjugate of a by g", namely gag^{-1}, is the same group element as the "right conjugate of a by g^{-1}", namely $a^{g^{-1}}$. Thus two elements or subsets of a group are "left conjugate" if and only if they are "right conjugate", and so the relation "conjugacy" is the same for the left and right corresponding actions. More generally, it is also an exercise (Exercise 1) to see that for any corresponding left and right actions the orbits are the same.

We have consistently used left actions since they are compatible with the notation of applying functions on the left (i.e. with the notation $\varphi(g)$); in this way left multiplication on the left cosets of a subgroup is a left action. Similarly, right multiplication on the right cosets of a subgroup is a right action and the associated permutation representation φ is a homomorphism provided the function $\varphi : G \to S_A$ is written on the right as $(g_1 g_2)\varphi$ (and also provided permutations in S_A are written on the right as functions from A to itself). There are instances where a set admits two actions by a group G: one naturally on the left and the other on the right, so that it is useful to be comfortable with both types of actions.

EXERCISES

Let G be a group.

1. Suppose G has a left action on a set A, denoted by $g \cdot a$ for all $g \in G$ and $a \in A$. Denote the corresponding right action on A by $a \cdot g$. Prove that the (equivalence) relations $\sim$ and $\sim'$ defined by

$$a \sim b \quad \text{if and only if} \quad a = g \cdot b \quad \text{for some } g \in G$$

and

$$a \sim' b \quad \text{if and only if} \quad a = b \cdot g \quad \text{for some } g \in G$$

are the same relation (i.e. $a \sim b$ if and only if $a \sim' b$).

2. Find all conjugacy classes and their sizes in the following groups:
(a) D_8 **(b)** Q_8 **(c)** A_4.
3. Find all the conjugacy classes and their sizes in the following groups:
(a) $Z_2 \times S_3$ **(b)** $S_3 \times S_3$ **(c)** $Z_3 \times A_4$.
4. Prove that if $S \subseteq G$ and $g \in G$ then $gN_G(S)g^{-1} = N_G(gSg^{-1})$ and $gC_G(S)g^{-1} = C_G(gSg^{-1})$.
5. If the center of G is of index n, prove that every conjugacy class has at most n elements.
6. Assume G is a non-abelian group of order 15. Prove that $Z(G) = 1$. Use the fact that $\langle g \rangle \leq C_G(g)$ for all $g \in G$ to show that there is at most one possible class equation for G. [Use Exercise 36, Section 3.1.]
7. For $n = 3, 4, 6$ and 7 make lists of the partitions of n and give representatives for the corresponding conjugacy classes of S_n.
8. Prove that $Z(S_n) = 1$ for all $n \geq 3$.
9. Show that $|C_{S_n}((1\,2)(3\,4))| = 8 \cdot (n-4)!$ for all $n \geq 4$. Determine the elements in this centralizer explicitly.
10. Let σ be the 5-cycle $(1\ 2\ 3\ 4\ 5)$ in S_5. In each of (a) to (c) find an explicit element $\tau \in S_5$ which accomplishes the specified conjugation:

(a) $\tau\sigma\tau^{-1} = \sigma^2$
(b) $\tau\sigma\tau^{-1} = \sigma^{-1}$
(c) $\tau\sigma\tau^{-1} = \sigma^{-2}$.

11. In each of (a) – (d) determine whether σ_1 and σ_2 are conjugate. If they are, give an explicit permutation τ such that $\tau\sigma_1\tau^{-1} = \sigma_2$
(a) $\sigma_1 = (1\ 2)(3\ 4\ 5)$ and $\sigma_2 = (1\ 2\ 3)(4\ 5)$
(b) $\sigma_1 = (1\ 5)(3\ 7\ 2)(10\ 6\ 8\ 11)$ and $\sigma_2 = (3\ 7\ 5\ 10)(4\ 9)(13\ 11\ 2)$
(c) $\sigma_1 = (1\ 5)(3\ 7\ 2)(10\ 6\ 8\ 11)$ and $\sigma_2 = \sigma_1^3$
(d) $\sigma_1 = (1\ 3)(2\ 4\ 6)$ and $\sigma_2 = (3\ 5)(2\ 4)(5\ 6)$.

12. Find a representative for each conjugacy class of elements of order 4 in S_8 and in S_{12}.

13. Find all finite groups which have exactly two conjugacy classes.

14. In Exercise 1 of Section 2 two labellings of the elements of the Klein 4-group $V = \{1, a, b, c\}$ were chosen to give two versions of the left regular representation of V into S_4. Let π_1 be the version of regular representation obtained in part (a) of that exercise and let π_2 be the version obtained via the labelling in part (b). Let $\tau = (2\ 4\ 3)$. Show that $\tau \circ \pi_1(g) \circ \tau^{-1} = \pi_2(g)$ for each $g \in V$ (i.e. conjugation by τ sends the image of π_1 to the image of π_2 elementwise).

15. Find an element of S_8 which conjugates the subgroup of S_8 obtained in part (a) of Exercise 3, Section 2 to the subgroup of S_8 obtained in part (b) of that same exercise (both of these subgroups are isomorphic to D_8).

16. Find an element of S_4 which conjugates the subgroup of S_4 obtained in part (a) of Exercise 5, Section 2 to the subgroup of S_4 obtained in part (b) of that same exercise (both of these subgroups are isomorphic to D_8).

17. Let A be a nonempty set and let X be any subset of S_A. Let

$$F(X) = \{a \in A \mid \sigma(a) = a \text{ for all } \sigma \in X\} \qquad \text{—the } \textit{fixed set} \text{ of } X.$$

Let $M(X) = A - F(X)$ be the elements which are *moved* by some element of X. Let $D = \{\sigma \in S_A \mid |M(\sigma)| < \infty\}$. Prove that D is a normal subgroup of S_A.

18. Let A be a set, let H be a subgroup of S_A and let $F(H)$ be the fixed points of H on A as defined in the preceding exercise. Prove that if $\tau \in N_{S_A}(H)$ then τ stabilizes the set $F(H)$ and its complement $A - F(H)$.

19. Assume H is a normal subgroup of G, $\mathcal{K}$ is a conjugacy class of G contained in H and $x \in \mathcal{K}$. Prove that $\mathcal{K}$ is a union of k conjugacy classes of equal size in H, where $k = |G : HC_G(x)|$. Deduce that a conjugacy class in S_n which consists of even permutations is either a single conjugacy class under the action of A_n or is a union of two classes of the same size in A_n. [Let $A = C_G(x)$ and $B = H$ so $A \cap B = C_H(x)$. Draw the lattice diagram associated to the Second Isomorphism Theorem and interpret the appropriate indices. See also Exercise 9, Section 1.]

20. Let $\sigma \in A_n$. Show that all elements in the conjugacy class of σ in S_n (i.e. all elements of the same cycle type as σ) are conjugate in A_n if and only if σ commutes with an odd permutation [Use the preceding exercise.]

21. Let $\mathcal{K}$ be a conjugacy class in S_n and assume that $\mathcal{K} \subseteq A_n$. Show $\sigma \in S_n$ does *not* commute with any odd permutation if and only if the cycle type of σ consists of distinct odd integers. Deduce that $\mathcal{K}$ is a (single) conjugacy class in A_n if and only if the cycle type of an element of $\mathcal{K}$ consists of distinct odd integers. [Assume first that $\sigma \in \mathcal{K}$ does not commute with any odd permutation. Observe that σ commutes with each individual cycle in its cycle decomposition —use this to show that all its cycles must be of odd length. If two cycles have the same odd length, k, find a product of k transpositions which interchanges them and commutes with

σ. Conversely, if the cycle decomposition of σ consists of distinct integers, prove that σ commutes *only* with the group generated by the cycles in its cycle decomposition.]

22. Show that if n is odd then the set of all n-cycles consists of two conjugacy classes of equal size in A_n.

23. Recall (cf. Exercise 16, Section 2.4) that a proper subgroup M of G is called *maximal* if whenever $M \leq H \leq G$, either $H = M$ or $H = G$. Prove that if M is a maximal subgroup of G then either $N_G(M) = M$ or $N_G(M) = G$. Deduce that if M is a maximal subgroup of G that is not normal in G then the number of nonidentity elements of G that are contained in conjugates of M is at most $(|M| - 1)|G : M|$.

24. Assume H is a proper subgroup of the finite group G. Prove $G \neq \cup_{g \in G} gHg^{-1}$, i.e. G is not the union of the conjugates of any proper subgroup. [Put H in some maximal subgroup and use the preceding exercise.]

25. Let $G = GL_2(\mathbb{C})$ and let $H = \{\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in \mathbb{C},\ ac \neq 0\}$. Prove that every element of G is conjugate to some element of the subgroup H and deduce that G is the union of conjugates of H. [Show that every element of $GL_2(\mathbb{C})$ has an eigenvector.]

26. Let G be a transitive permutation group on the finite set A with $|A| > 1$. Show that there is some $\sigma \in G$ such that $\sigma(a) \neq a$ for all $a \in A$ (such an element σ is called *fixed point free*).

27. Let $g_1, g_2, \ldots, g_r$ be representatives of the conjugacy classes of the finite group G and assume these elements pairwise commute. Prove that G is abelian.

28. Let p and q be primes with $p < q$. Prove that a non-abelian group G of order pq has a nonnormal subgroup of index q, so that there exists an injective homomorphism into S_q. Deduce that G is isomorphic to a subgroup of the normalizer in S_q of the cyclic group generated by the q-cycle $(1\,2 \ldots q)$.

29. Let p be a prime and let G be a group of order p^α. Prove that G has a subgroup of order p^β, for every β with $0 \leq \beta \leq \alpha$. [Use Theorem 8 and induction on α.]

30. If G is a group of odd order, prove for any nonidentity element $x \in G$ that x and x^{-1} are not conjugate in G.

31. Using the usual generators and relations for the dihedral group D_{2n} (cf. Section 1.2) show that for $n = 2k$ an even integer the conjugacy classes in D_{2n} are: $\{1\}$, $\{r^k\}$, $\{r^{\pm 1}\}$, $\{r^{\pm 2}\}$, $\ldots$, $\{r^{\pm(k-1)}\}$, $\{sr^{2b} \mid b = 1, \ldots, k\}$ and $\{sr^{2b-1} \mid b = 1, \ldots, k\}$. Give the class equation for D_{2n}.

32. For $n = 2k + 1$ an odd integer show that the conjugacy classes in D_{2n} are: $\{1\}$, $\{r^{\pm 1}\}$, $\{r^{\pm 2}\}$, $\ldots$, $\{r^{\pm k}\}$, $\{sr^b \mid b = 1, \ldots, n\}$. Give the class equation for D_{2n}.

33. This exercise gives a formula for the size of each conjugacy class in S_n. Let σ be a permutation in S_n and let $m_1, m_2, \ldots, m_s$ be the *distinct* integers which appear in the cycle type of σ (including 1-cycles). For each $i \in \{1, 2, \ldots, s\}$ assume σ has k_i cycles of length m_i (so that $\Sigma_{i=1}^s k_i m_i = n$). Prove that the number of conjugates of σ is:

$$\frac{n!}{(k_1! m_1^{k_1})(k_2! m_2^{k_2}) \ldots (k_s! m_s^{k_s})}.$$

[See Exercises 16 and 17 in Section 1.3 where this formula was given in some special cases.]

34. Prove that if p is a prime and P is a subgroup of S_p of order p, then $|N_{S_p}(P)| = p(p-1)$. [Argue that every conjugate of P contains exactly $p - 1$ p-cycles and use the formula for the number of p-cycles to compute the index of $N_{S_p}(P)$ in S_p.]

35. Let p be a prime. Find a formula for the number of conjugacy classes of elements of order p in S_n (using the greatest integer function).

4.4 AUTOMORPHISMS

Definition. Let G be a group. An isomorphism from G onto itself is called an *automorphism* of G. The set of all automorphisms of G is denoted by $\mathrm{Aut}(G)$.

We leave as an exercise the simple verification that $\mathrm{Aut}(G)$ is a group under composition of automorphisms, the *automorphism group* of G (composition of automorphisms is defined since the domain and range of each automorphism is the same). Notice that automorphisms of a group G are, in particular, permutations of the set G so $\mathrm{Aut}(G)$ is a subgroup of S_G.

One of the most important examples of an automorphism of a group G is provided by conjugation by a fixed element in G. The next result discusses this in a slightly more general context.

Proposition 13. Let H be a normal subgroup of the group G. Then G acts by conjugation on H as automorphisms of H. More specifically, the action of G on H by conjugation is defined for each $g \in G$ by

$$h \mapsto ghg^{-1} \qquad \text{for each } h \in H.$$

For each $g \in G$, conjugation by g is an automorphism of H. The permutation representation afforded by this action is a homomorphism of G into $\mathrm{Aut}(H)$ with kernel $C_G(H)$. In particular, $G/C_G(H)$ is isomorphic to a subgroup of $\mathrm{Aut}(H)$.

Proof: (cf. Exercise 17, Section 1.7) Let φ_g be conjugation by g. Note that because g normalizes H, φ_g maps H to itself. Since we have already seen that conjugation defines an action, it follows that $\varphi_1 = 1$ (the identity map on H) and $\varphi_a \circ \varphi_b = \varphi_{ab}$ for all $a, b \in G$. Thus each φ_g gives a bijection from H to itself since it has a 2-sided inverse $\varphi_{g^{-1}}$. Each φ_g is a homomorphism from H to H because

$$\varphi_g(hk) = g(hk)g^{-1} = gh(gg^{-1})kg^{-1} = (ghg^{-1})(gkg^{-1}) = \varphi_g(h)\varphi_g(k)$$

for all $h, k \in H$. This proves that conjugation by any fixed element of G defines an automorphism of H.

By the preceding remark, the permutation representation $\psi : G \to S_H$ defined by $\psi(g) = \varphi_g$ (which we have already proved is a homomorphism) has image contained in the subgroup $\mathrm{Aut}(H)$ of S_H. Finally,

$$\begin{aligned} \ker \psi &= \{g \in G \mid \varphi_g = \mathrm{id}\} \\ &= \{g \in G \mid ghg^{-1} = h \text{ for all } h \in H\} \\ &= C_G(H). \end{aligned}$$

The First Isomorphism Theorem implies the final statement of the proposition.

Proposition 13 shows that a group acts by conjugation on a normal subgroup as *structure preserving* permutations, i.e. as automorphisms. In particular, this action must send subgroups to subgroups, elements of order n to elements of order n, etc. Two specific applications of this proposition are described in the next two corollaries.

Corollary 14. If K is any subgroup of the group G and $g \in G$, then $K \cong gKg^{-1}$. Conjugate elements and conjugate subgroups have the same order.

Proof: Letting $G = H$ in the proposition shows that conjugation by $g \in G$ is an automorphism of G, from which the corollary follows.

Corollary 15. For any subgroup H of a group G, the quotient group $N_G(H)/C_G(H)$ is isomorphic to a subgroup of $\mathrm{Aut}(H)$. In particular, $G/Z(G)$ is isomorphic to a subgroup of $\mathrm{Aut}(G)$.

Proof: Since H is a normal subgroup of the group $N_G(H)$, Proposition 13 (applied with $N_G(H)$ playing the role of G) implies the first assertion. The second assertion is the special case when $H = G$, in which case $N_G(G) = G$ and $C_G(G) = Z(G)$.

Definition. Let G be a group and let $g \in G$. Conjugation by g is called an *inner automorphism* of G and the subgroup of $\mathrm{Aut}(G)$ consisting of all inner automorphisms is denoted by $\mathrm{Inn}(G)$.

Note that the collection of inner automorphisms of G is in fact a subgroup of $\mathrm{Aut}(G)$ and that by Corollary 15, $\mathrm{Inn}(G) \cong G/Z(G)$. Note also that if H is a normal subgroup of G, conjugation by an element of G when restricted to H is an automorphism of H but need not be an inner automorphism of H (as we shall see).

Examples

(1) A group G is abelian if and only if every inner automorphism is trivial. If H is an abelian normal subgroup of G and H is not contained in $Z(G)$, then there is some $g \in G$ such that conjugation by g restricted to H is not an inner automorphism of H. An explicit example of this is $G = A_4$, H is the Klein 4-group in G and g is any 3-cycle.
(2) Since $Z(Q_8) = \langle -1 \rangle$ we have $\mathrm{Inn}(Q_8) \cong V_4$.
(3) Since $Z(D_8) = \langle r^2 \rangle$ we have $\mathrm{Inn}(D_8) \cong V_4$.
(4) Since for all $n \geq 3$, $Z(S_n) = 1$ we have $\mathrm{Inn}(S_n) \cong S_n$.

Corollary 15 shows that any information we have about the automorphism group of a subgroup H of a group G translates into information about $N_G(H)/C_G(H)$. For example, if $H \cong Z_2$, then since H has unique elements of orders 1 and 2, Corollary 14 forces $\mathrm{Aut}(H) = 1$. Thus if $H \cong Z_2$, $N_G(H) = C_G(H)$; if in addition H is a normal subgroup of G, then $H \leq Z(G)$ (cf. Exercise 10, Section 2.2).

Although the preceding example was fairly trivial, it illustrates that the action of G by conjugation on a *normal* subgroup H can be restricted by knowledge of the automorphism group of H. This in turn can be used to investigate the structure of G and will lead to some classification theorems when we consider semidirect products in Section 5.5.

A notion which will be used in later sections most naturally warrants introduction here:

Definition. A subgroup H of a group G is called *characteristic* in G, denoted $H \operatorname{char} G$, if every automorphism of G maps H to itself, i.e. $\sigma(H) = H$ for all $\sigma \in \mathrm{Aut}(G)$.

Results concerning characteristic subgroups which we shall use later (and whose proofs are relegated to the exercises) are:

(1) characteristic subgroups are normal,
(2) if H is the unique subgroup of G of a given order, then H is characteristic in G, and
(3) if K char H and $H \trianglelefteq G$, then $K \trianglelefteq G$ (so although "normality" is not a transitive property (i.e. a normal subgroup of a normal subgroup need not be normal), a characteristic subgroup of a normal subgroup is normal).

Thus we may think of characteristic subgroups as "strongly normal" subgroups. For example, property (2) and Theorem 2.7 imply that every subgroup of a cyclic group is characteristic.

We close this section with some results on automorphism groups of specific groups.

Proposition 16. The automorphism group of the cyclic group of order n is isomorphic to $(\mathbb{Z}/n\mathbb{Z})^\times$, an abelian group of order $\varphi(n)$ (where φ is Euler's function).

Proof: Let x be a generator of the cyclic group Z_n. If $\psi \in \mathrm{Aut}(Z_n)$, then $\psi(x) = x^a$ for some $a \in \mathbb{Z}$ and the integer a uniquely determines ψ. Denote this automorphism by ψ_a. As usual, since $|x| = n$, the integer a is only defined mod n. Since φ is an automorphism, x and x^a must have the same order, hence $(a, n) = 1$. Furthermore, for every a relatively prime to n, the map $x \mapsto x^a$ is an automorphism of Z_n. Hence we have a surjective map

$$\begin{aligned} \Psi : \mathrm{Aut}(Z_n) &\to (\mathbb{Z}/n\mathbb{Z})^\times \\ \psi_a &\mapsto a \pmod{n}. \end{aligned}$$

The map Ψ is a homomorphism because

$$\psi_a \circ \psi_b(x) = \psi_a(x^b) = (x^b)^a = x^{ab} = \psi_{ab}(x)$$

for all $\psi_a, \psi_b \in \mathrm{Aut}(Z_n)$, so that

$$\Psi(\psi_a \circ \psi_b) = \Psi(\psi_{ab}) = ab \pmod{n} = \Psi(\psi_a)\Psi(\psi_b).$$

Finally, Ψ is clearly injective, hence is an isomorphism.

A complete description of the isomorphism type of $\mathrm{Aut}(Z_n)$ is given at the end of Section 9.5.

Example

Assume G is a group of order pq, where p and q are primes (not necessarily distinct) with $p \leq q$. If $p \nmid q - 1$, we prove G is abelian.

If $Z(G) \neq 1$, Lagrange's Theorem forces $G/Z(G)$ to be cyclic, hence G is abelian by Exercise 36, Section 3.1. Hence we may assume $Z(G) = 1$.

If every nonidentity element of G has order p, then the centralizer of every non-

identity element has index q, so the class equation for G reads

$$pq = 1 + kq.$$

This is impossible since q divides pq and kq but not 1. Thus G contains an element, x, of order q.

Let $H = \langle x \rangle$. Since H has index p and p is the smallest prime dividing $|G|$, the subgroup H is normal in G by Corollary 5. Since $Z(G) = 1$, we must have $C_G(H) = H$. Thus $G/H = N_G(H)/C_G(H)$ is a group of order p isomorphic to a subgroup of $\mathrm{Aut}(H)$ by Corollary 15. But by Proposition 16, $\mathrm{Aut}(H)$ has order $\varphi(q) = q - 1$, which by Lagrange's Theorem would imply $p \mid q - 1$, contrary to assumption. This shows that G must be abelian.

One can check that every group of order pq, where p and q are distinct primes with $p < q$ and $p \nmid q - 1$ is *cyclic* (see the exercises). This is the first instance where there is a unique isomorphism type of group whose order is *composite*. For instance, every group of order 15 is cyclic.

The next proposition summarizes some results on automorphism groups of known groups and will be proved later. Part 3 of this proposition illustrates how the theory of vector spaces comes into play in group theory.

Proposition 17.

(1) If p is an odd prime and $n \in \mathbb{Z}^+$, then the automorphism group of the cyclic group of order p is cyclic of order $p - 1$. More generally, the automorphism group of the cyclic group of order p^n is cyclic of order $p^{n-1}(p-1)$ (cf. Corollary 20, Section 9.5).

(2) For all $n \geq 3$ the automorphism group of the cyclic group of order 2^n is isomorphic to $Z_2 \times Z_{2^{n-2}}$, and in particular is not cyclic but has a cyclic subgroup of index 2 (cf. Corollary 20, Section 9.5).

(3) Let p be a prime and let V be an abelian group (written additively) with the property that $pv = 0$ for all $v \in V$. If $|V| = p^n$, then V is an n-dimensional vector space over the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. The automorphisms of V are precisely the nonsingular linear transformations from V to itself, that is

$$\mathrm{Aut}(V) \cong GL(V) \cong GL_n(\mathbb{F}_p).$$

In particular, the order of $\mathrm{Aut}(V)$ is as given in Section 1.4 (cf. the examples in Section 10.2 and 11.1).

(4) For all $n \neq 6$ we have $\mathrm{Aut}(S_n) = \mathrm{Inn}(S_n) \cong S_n$ (cf. Exercise 18). For $n = 6$ we have $|\mathrm{Aut}(S_6) : \mathrm{Inn}(S_6)| = 2$ (cf. the following Exercise 19 and also Exercise 10 in Section 6.3).

(5) $\mathrm{Aut}(D_8) \cong D_8$ and $\mathrm{Aut}(Q_8) \cong S_4$ (cf. the following Exercises 4 and 5 and also Exercise 9 in Section 6.3).

The group V described in Part 3 of the proposition is called the *elementary abelian* group of order p^n (we shall see in Chapter 5 that it is uniquely determined

up to isomorphism by p and n). The Klein 4-group, V_4, is the elementary abelian group of order 4. This proposition asserts that

$$\mathrm{Aut}(V_4) \cong GL_2(\mathbb{F}_2).$$

By the exercises in Section 1.4, the latter group has order 6. But $\mathrm{Aut}(V_4)$ permutes the 3 nonidentity elements of V_4, and this action of $\mathrm{Aut}(V_4)$ on $V_4 - \{1\}$ gives an injective permutation representation of $\mathrm{Aut}(V_4)$ into S_3. By order considerations, the homomorphism is onto, so

$$\mathrm{Aut}(V_4) \cong GL_2(\mathbb{F}_2) \cong S_3.$$

Note that V_4 is abelian, so $\mathrm{Inn}(V_4) = 1$.

For any prime p, the elementary abelian group of order p^2 is $Z_p \times Z_p$. Its automorphism group, $GL_2(\mathbb{F}_p)$, has order $p(p-1)(p+1)$. Thus Corollary 9 implies that for p a prime

$$\text{if } |P| = p^2, \quad |\mathrm{Aut}(P)| = p(p-1) \text{ or } p(p-1)(p+1)$$

according to whether P is cyclic or elementary abelian, respectively.

Example

Suppose G is a group of order $45 = 3^2 5$ with a normal subgroup P of order 3^2. We show that G is necessarily abelian.

The quotient $G/C_G(P)$ is isomorphic to a subgroup of $\mathrm{Aut}(P)$ by Corollary 15, and $\mathrm{Aut}(P)$ has order 6 or 24 (according to whether P is cyclic or elementary abelian, respectively) by the preceding paragraph. On the other hand, since the order of P is the square of a prime, P is an abelian group, hence $P \leq C_G(P)$. It follows that $|C_G(P)|$ is divisible by 9, which implies $|G/C_G(P)|$ is 1 or 5. Together these imply $|G/C_G(P)| = 1$, i.e. $C_G(P) = G$ and $P \leq Z(G)$. Since then $G/Z(G)$ is cyclic, G must be an abelian group.

EXERCISES

Let G be a group.

1. Prove that if $\sigma \in \mathrm{Aut}(G)$ and φ_g is conjugation by g then $\sigma\varphi_g\sigma^{-1} = \varphi_{\sigma(g)}$. Deduce that $\mathrm{Inn}(G) \trianglelefteq \mathrm{Aut}(G)$. (The group $\mathrm{Aut}(G)/\mathrm{Inn}(G)$ is called the *outer automorphism group* of G.)
2. Prove that if G is an abelian group of order pq, where p and q are distinct primes, then G is cyclic. [Use Cauchy's Theorem to produce elements of order p and q and consider the order of their product.]
3. Prove that under any automorphism of D_8, r has at most 2 possible images and s has at most 4 possible images (r and s are the usual generators — cf. Section 1.2). Deduce that $|\mathrm{Aut}(D_8)| \leq 8$.
4. Use arguments similar to those in the preceding exercise to show $|\mathrm{Aut}(Q_8)| \leq 24$.
5. Use the fact that $D_8 \trianglelefteq D_{16}$ to prove that $\mathrm{Aut}(D_8) \cong D_8$.
6. Prove that characteristic subgroups are normal. Give an example of a normal subgroup that is not characteristic.
7. Prove that if H is the unique subgroup of a given order in a group G then H is characteristic in G.

8. Let G be a group with subgroups H and K with $H \leq K$.
   (a) Prove that if H is characteristic in K and K is normal in G then H is normal in G.
   (b) Prove that if H is characteristic in K and K is characteristic in G then H is characteristic in G. Use this to prove that the Klein 4-group V_4 is characteristic in S_4.
   (c) Give an example to show that if H is normal in K and K is characteristic in G then H need not be normal in G.
9. If r, s are the usual generators for the dihedral group D_{2n}, use the preceding two exercises to deduce that every subgroup of $\langle r \rangle$ is normal in D_{2n}.
10. Let G be a group, let A be an abelian normal subgroup of G, and write $\overline{G} = G/A$. Show that $\overline{G}$ acts (on the left) by conjugation on A by $\overline{g} \cdot a = gag^{-1}$, where g is *any* representative of the coset $\overline{g}$ (in particular, show that this action is well defined). Give an explicit example to show that this action is not well defined if A is non-abelian.
11. Prove that if p is a prime and P is a subgroup of S_p of order p, then $N_{S_p}(P)/C_{S_p}(P) \cong \mathrm{Aut}(P)$. [Use Exercise 34, Section 3.]
12. Let G be a group of order 3825. Prove that if H is a normal subgroup of order 17 in G then $H \leq Z(G)$.
13. Let G be a group of order 203. Prove that if H is a normal subgroup of order 7 in G then $H \leq Z(G)$. Deduce that G is abelian in this case.
14. Let G be a group of order 1575. Prove that if H is a normal subgroup of order 9 in G then $H \leq Z(G)$.
15. Prove that each of the following (multiplicative) groups is cyclic: $(\mathbb{Z}/5\mathbb{Z})^\times$, $(\mathbb{Z}/9\mathbb{Z})^\times$ and $(\mathbb{Z}/18\mathbb{Z})^\times$.
16. Prove that $(\mathbb{Z}/24\mathbb{Z})^\times$ is an elementary abelian group of order 8. (We shall see later that $(\mathbb{Z}/n\mathbb{Z})^\times$ is an elementary abelian group if and only if $n \mid 24$.)
17. Let $G = \langle x \rangle$ be a cyclic group of order n. For each of the following values of n, write out the elements of $\mathrm{Aut}(G)$ explicitly (by Proposition 16 above we know $\mathrm{Aut}(G) \cong (\mathbb{Z}/n\mathbb{Z})^\times$, so for each element $a \in (\mathbb{Z}/n\mathbb{Z})^\times$, write out explicitly what the automorphism ψ_a does to the elements $\{1, x, x^2, \ldots, x^{n-1}\}$ of G):
   (a) $n = 2$
   (b) $n = 3$
   (c) $n = 4$
   (d) $n = 5$
   (e) $n = 6$.
18. This exercise shows that for $n \neq 6$ every automorphism of S_n is inner. Fix an integer $n \geq 2$ with $n \neq 6$.
   (a) Prove that the automorphism group of a group G permutes the conjugacy classes of G, i.e. for each $\sigma \in \mathrm{Aut}(G)$ and each conjugacy class $\mathcal{K}$ of G the set $\sigma(\mathcal{K})$ is also a conjugacy class of G.
   (b) Let $\mathcal{K}$ be the conjugacy class of transpositions in S_n and let $\mathcal{K}'$ be the conjugacy class of any element of order 2 in S_n that is not a transposition. Prove that $|\mathcal{K}| \neq |\mathcal{K}'|$. Deduce that any automorphism of S_n sends transpositions to transpositions. [See Exercise 33 in Section 3.]
   (c) Prove that for each $\sigma \in \mathrm{Aut}(S_n)$

$$\sigma : (1\ 2) \mapsto (a\ b_2), \qquad \sigma : (1\ 3) \mapsto (a\ b_3), \quad \ldots, \quad \sigma : (1\ n) \mapsto (a\ b_n)$$

   for some distinct integers $a, b_2, b_3, \ldots, b_n \in \{1, 2, \ldots, n\}$.
   (d) Show that $(1\ 2)$, $(1\ 3)$, $\ldots$, $(1\ n)$ generate S_n and deduce that any automorphism of

S_n is uniquely determined by its action on these elements. Use (c) to show that S_n has at most $n!$ automorphisms and conclude that $\mathrm{Aut}(S_n) = \mathrm{Inn}(S_n)$ for $n \neq 6$.

19. This exercise shows that $|\mathrm{Aut}(S_6) : \mathrm{Inn}(S_6)| \leq 2$ (Exercise 10 in Section 6.3 shows that equality holds by exhibiting an automorphism of S_6 that is not inner).

(a) Let $\mathcal{K}$ be the conjugacy class of transpositions in S_6 and let $\mathcal{K}'$ be the conjugacy class of any element of order 2 in S_6 that is not a transposition. Prove that $|\mathcal{K}| \neq |\mathcal{K}'|$ unless $\mathcal{K}'$ is the conjugacy class of products of three disjoint transpositions. Deduce that $\mathrm{Aut}(S_6)$ has a subgroup of index at most 2 which sends transpositions to transpositions.

(b) Prove that $|\mathrm{Aut}(S_6) : \mathrm{Inn}(S_6)| \leq 2$. [Follow the same steps as in (c) and (d) of the preceding exercise to show that any automorphism that sends transpositions to transpositions is inner.]

The next exercise introduces a subgroup, $J(P)$, which (like the center of P) is defined for an arbitrary finite group P (although in most applications P is a group whose order is a power of a prime). This subgroup was defined by J. Thompson in 1964 and it now plays a pivotal role in the study of finite groups, in particular, in the classification of finite simple groups.

20. For any finite group P let $d(P)$ be the minimum number of generators of P (so, for example, $d(P) = 1$ if and only if P is a nontrivial cyclic group and $d(Q_8) = 2$). Let $m(P)$ be the maximum of the integers $d(A)$ as A runs over all *abelian* subgroups of P (so, for example, $m(Q_8) = 1$ and $m(D_8) = 2$). Define

$$J(P) = \langle A \mid A \text{ is an abelian subgroup of } P \text{ with } d(A) = m(P) \rangle$$

$J(P)$ is called the *Thompson subgroup* of P.

(a) Prove that $J(P)$ is a characteristic subgroup of P.

(b) For each of the following groups P list all abelian subgroups A of P that satisfy $d(A) = m(P)$: Q_8, D_8, D_{16} and QD_{16} (where QD_{16} is the quasidihedral group of order 16 defined in Exercise 11 of Section 2.5). [Use the lattices of subgroups for these groups in Section 2.5.]

(c) Show that $J(Q_8) = Q_8$, $J(D_8) = D_8$, $J(D_{16}) = D_{16}$ and $J(QD_{16})$ is a dihedral subgroup of order 8 in QD_{16}.

(d) Prove that if $Q \leq P$ and $J(P)$ is a subgroup of Q, then $J(P) = J(Q)$. Deduce that if P is a subgroup (not necessarily normal) of the finite group G and $J(P)$ is contained in some subgroup Q of P such that $Q \trianglelefteq G$, then $J(P) \trianglelefteq G$.

4.5 SYLOW'S THEOREM

In this section we prove a partial converse to Lagrange's Theorem and derive numerous consequences, some of which will lead to classification theorems in the next chapter.

Definition. Let G be a group and let p be a prime.

(1) A group of order p^α for some $\alpha \geq 1$ is called a *p-group*. Subgroups of G which are p-groups are called *p-subgroups*.

(2) If G is a group of order $p^\alpha m$, where $p \nmid m$, then a subgroup of order p^α is called a *Sylow p-subgroup* of G.

(3) The set of Sylow p-subgroups of G will be denoted by $Syl_p(G)$ and the number of Sylow p-subgroups of G will be denoted by $n_p(G)$ (or just n_p when G is clear from the context).

Theorem 18. (Sylow's Theorem) Let G be a group of order $p^\alpha m$, where p is a prime not dividing m.

(1) Sylow p-subgroups of G exist, i.e. $Syl_p(G) \neq \emptyset$.

(2) If P is a Sylow p-subgroup of G and Q is any p-subgroup of G, then there exists $g \in G$ such that $Q \leq gPg^{-1}$, i.e. Q is contained in some conjugate of P. In particular, any two Sylow p-subgroups of G are conjugate in G.

(3) The number of Sylow p-subgroups of G is of the form $1 + kp$, i.e.

$$n_p \equiv 1 (\text{mod } p).$$

Further, n_p is the index in G of the normalizer $N_G(P)$ for any Sylow p-subgroup P, hence

$$n_p \mid m.$$

We first prove the following lemma:

Lemma 19. Let $P \in Syl_p(G)$. If Q is any p-subgroup of G, then $Q \cap N_G(P) = Q \cap P$.

Proof: Let $H = N_G(P) \cap Q$. Since $P \leq N_G(P)$ it is clear that $P \cap Q \leq H$, so we must prove the reverse inclusion. Since by definition $H \leq Q$, this is equivalent to showing $H \leq P$. We do this by demonstrating that PH is a p-subgroup of G containing both P and H; but P is a p-subgroup of G of largest possible order, so we must have $PH = P$, i.e. $H \leq P$.

Since $H \leq N_G(P)$, by Corollary 15 in Section 3.2, PH is a subgroup. By Proposition 13 in the same section

$$|PH| = \frac{|P||H|}{|P \cap H|}.$$

All the numbers in the above quotient are powers of p, so PH is a p-group. Moreover, P is a subgroup of PH so the order of PH is divisible by p^α, the largest power of p which divides $|G|$. These two facts force $|PH| = p^\alpha = |P|$. This in turn implies $P = PH$ and $H \leq P$. This establishes the lemma.

Proof of Sylow's Theorem: (1) Proceed by induction on $|G|$. If $|G| = 1$, there is nothing to prove. Assume inductively the existence of Sylow p-subgroups for all groups of order less than $|G|$.

If p divides $|Z(G)|$, then by Cauchy's Theorem for abelian groups (Proposition 21, Section 3.4) $Z(G)$ has a subgroup, N, of order p. Let $\overline{G} = G/N$, so that $|\overline{G}| = p^{\alpha-1}m$. By induction, $\overline{G}$ has a subgroup $\overline{P}$ of order $p^{\alpha-1}$. If we let P be the subgroup of G containing N such that $P/N = \overline{P}$ then $|P| = |P/N| \cdot |N| = p^\alpha$ and P is a Sylow p-subgroup of G. We are reduced to the case when p does not divide $|Z(G)|$.

Let $g_1, g_2, \ldots, g_r$ be representatives of the distinct non-central conjugacy classes

of G. The class equation for G is

$$|G| = |Z(G)| + \sum_{i=1}^{r} |G : C_G(g_i)|.$$

If $p \mid |G : C_G(g_i)|$ for all i, then since $p \mid |G|$, we would also have $p \mid |Z(G)|$, a contradiction. Thus for some i, p does not divide $|G : C_G(g_i)|$. For this i let $H = C_G(g_i)$ so that

$$|H| = p^\alpha k, \quad \text{where } p \nmid k.$$

Since $g_i \notin Z(G)$, $|H| < |G|$. By induction, H has a Sylow p-subgroup, P, which of course is also a subgroup of G. Since $|P| = p^\alpha$, P is a Sylow p-subgroup of G. This completes the induction and establishes (1).

Before proving (2) and (3) we make some calculations. By (1) there exists a Sylow p-subgroup, P, of G. Let

$$\{P_1, P_2, \ldots, P_r\} = \mathcal{S}$$

be the set of all conjugates of P (i.e. $\mathcal{S} = \{gPg^{-1} \mid g \in G\}$) and let Q be *any* p-subgroup of G. By definition of $\mathcal{S}$, G, hence also Q, acts by conjugation on $\mathcal{S}$. Write $\mathcal{S}$ as a disjoint union of orbits under this action by Q:

$$\mathcal{S} = \mathcal{O}_1 \cup \mathcal{O}_2 \cup \cdots \cup \mathcal{O}_s$$

where $r = |\mathcal{O}_1| + \cdots + |\mathcal{O}_s|$. Keep in mind that r does not depend on Q but the number of Q-orbits s does (note that by definition, G has only one orbit on $\mathcal{S}$ but a subgroup Q of G may have more than one orbit). Renumber the elements of $\mathcal{S}$ if necessary so that the first s elements of $\mathcal{S}$ are representatives of the Q-orbits: $P_i \in \mathcal{O}_i$, $1 \le i \le s$. It follows from Proposition 2 that $|\mathcal{O}_i| = |Q : N_Q(P_i)|$. By definition, $N_Q(P_i) = N_G(P_i) \cap Q$ and by Lemma 19, $N_G(P_i) \cap Q = P_i \cap Q$. Combining these two facts gives:

$$|\mathcal{O}_i| = |Q : P_i \cap Q|, \qquad 1 \le i \le s. \tag{$*$}$$

We are now in a position to prove that $r \equiv 1 (\text{mod } p)$. Since Q was arbitrary we may take $Q = P_1$ above, so that $(*)$ gives

$$|\mathcal{O}_1| = 1.$$

Now, for all $i > 1$, $P_1 \neq P_i$, so $P_1 \cap P_i < P_1$. By $(*)$

$$|\mathcal{O}_i| = |P_1 : P_1 \cap P_i| > 1, \qquad 2 \le i \le s.$$

Since P_1 is a p-group, $|P_1 : P_1 \cap P_i|$ must be a power of p, so that

$$p \mid |\mathcal{O}_i|, \qquad 2 \le i \le s.$$

Thus

$$r = |\mathcal{O}_1| + (|\mathcal{O}_2| + \ldots + |\mathcal{O}_s|) \equiv 1 (\text{mod } p).$$

We now prove parts (2) and (3). Let Q be any p-subgroup of G. Suppose Q is not contained in P_i for any $i \in \{1, 2, \ldots, r\}$ (i.e. $Q \not\le gPg^{-1}$ for any $g \in G$). In this

situation, $Q \cap P_i < Q$ for all i, so by $(*)$

$$|\mathcal{O}_i| = |Q : Q \cap P_i| > 1, \qquad 1 \le i \le s.$$

Thus $p \mid |\mathcal{O}_i|$ for all i, so p divides $|\mathcal{O}_1| + \ldots + |\mathcal{O}_s| = r$. This contradicts the fact that $r \equiv 1 (\text{mod } p)$ (remember, r does not depend on the choice of Q). This contradiction proves $Q \le gPg^{-1}$ for some $g \in G$.

To see that all Sylow p-subgroups of G are conjugate, let Q be any Sylow p-subgroup of G. By the preceding argument, $Q \le gPg^{-1}$ for some $g \in G$. Since $|gPg^{-1}| = |Q| = p^\alpha$, we must have $gPg^{-1} = Q$. This establishes part (2) of the theorem. In particular, $\mathcal{S} = Syl_p(G)$ since *every* Sylow p-subgroup of G is conjugate to P, and so $n_p = r \equiv 1 (\text{mod } p)$, which is the first part of (3).

Finally, since all Sylow p-subgroups are conjugate, Proposition 6 shows that

$$n_p = |G : N_G(P)| \quad \text{for any } P \in Syl_p(G),$$

completing the proof of Sylow's Theorem.

Note that the conjugacy part of Sylow's Theorem together with Corollary 14 shows that *any two Sylow p-subgroups of a group (for the same prime p) are isomorphic*.

Corollary 20. Let P be a Sylow p-subgroup of G. Then the following are equivalent:

(1) P is the unique Sylow p-subgroup of G, i.e. $n_p = 1$

(2) P is normal in G

(3) P is characteristic in G

(4) All subgroups generated by elements of p-power order are p-groups, i.e. if X is any subset of G such that $|x|$ is a power of p for all $x \in X$, then $\langle X \rangle$ is a p-group.

Proof: If (1) holds, then $gPg^{-1} = P$ for all $g \in G$ since $gPg^{-1} \in Syl_p(G)$, i.e. P is normal in G. Hence (1) implies (2). Conversely, if $P \trianglelefteq G$ and $Q \in Syl_p(G)$, then by Sylow's Theorem there exists $g \in G$ such that $Q = gPg^{-1} = P$. Thus $Syl_p(G) = \{P\}$ and (2) implies (1).

Since characteristic subgroups are normal, (3) implies (2). Conversely, if $P \trianglelefteq G$, we just proved P is the unique subgroup of G of order p^α, hence $P \operatorname{char} G$. Thus (2) and (3) are equivalent.

Finally, assume (1) holds and suppose X is a subset of G such that $|x|$ is a power of p for all $x \in X$. By the conjugacy part of Sylow's Theorem, for each $x \in X$ there is some $g \in G$ such that $x \in gPg^{-1} = P$. Thus $X \subseteq P$, and so $\langle X \rangle \le P$, and $\langle X \rangle$ is a p-group. Conversely, if (4) holds, let X be the union of all Sylow p-subgroups of G. If P is any Sylow p-subgroup, P is a subgroup of the p-group $\langle X \rangle$. Since P is a p-subgroup of G of maximal order, we must have $P = \langle X \rangle$, so (1) holds.

Examples

Let G be a finite group and let p be a prime.

(1) If p does not divide the order of G, the Sylow p-subgroup of G is the trivial group (and all parts of Sylow's Theorem hold trivially). If $|G| = p^\alpha$, G is the unique Sylow p-subgroup of G.

(2) A finite abelian group has a unique Sylow p-subgroup for each prime p. This subgroup consists of all elements x whose order is a power of p. This is sometimes called the *p-primary component* of the abelian group.
(3) S_3 has three Sylow 2-subgroups: $\langle(1\,2)\rangle$, $\langle(2\,3)\rangle$ and $\langle(1\,3)\rangle$. It has a unique (hence normal) Sylow 3-subgroup: $\langle(1\,2\,3)\rangle = A_3$. Note that $3 \equiv 1 (\text{mod } 2)$.
(4) A_4 has a unique Sylow 2-subgroup: $\langle(1\,2)(3\,4), (1\,3)(2\,4)\rangle \cong V_4$. It has four Sylow 3-subgroups: $\langle(1\,2\,3)\rangle$, $\langle(1\,2\,4)\rangle$, $\langle(1\,3\,4)\rangle$ and $\langle(2\,3\,4)\rangle$. Note that $4 \equiv 1 (\text{mod } 3)$.
(5) S_4 has $n_2 = 3$ and $n_3 = 4$. Since S_4 contains a subgroup isomorphic to D_8, every Sylow 2-subgroup of S_4 is isomorphic to D_8.

Applications of Sylow's Theorem

We now give some applications of Sylow's Theorem. Most of the examples use Sylow's Theorem to prove that a group of a particular order is not simple. After discussing methods of constructing larger groups from smaller ones (for example, the formation of semidirect products) we shall be able to use these results to classify groups of some specific orders n (as we already did for $n = 15$).

Since Sylow's Theorem ensures the existence of p-subgroups of a finite group, it is worthwhile to study groups of prime power order more closely. This will be done in Chapter 6 and many more applications of Sylow's Theorem will be discussed there.

For groups of small order, the congruence condition of Sylow's Theorem alone is often sufficient to force the existence of a *normal* subgroup. The first step in any numerical application of Sylow's Theorem is to factor the group order into prime powers. The largest prime divisors of the group order tend to give the fewest possible values for n_p (for example, the congruence condition on n_2 gives no restriction whatsoever), which limits the structure of the group G. In the following examples we shall see situations where Sylow's Theorem alone does not force the existence of a normal subgroup, however some additional argument (often involving studying the elements of order p for a number of different primes p) proves the existence of a normal Sylow subgroup.

Example: (Groups of order *pq*, *p*, and *q* primes with p < q)

Suppose $|G| = pq$ for primes p and q with $p < q$. Let $P \in Syl_p(G)$ and let $Q \in Syl_q(G)$. We show that Q is normal in G and if P is also normal in G, then G is cyclic.

Now the three conditions: $n_q = 1 + kq$ for some $k \geq 0$, n_q divides p and $p < q$, together force $k = 0$. Since $n_q = 1$, $Q \trianglelefteq G$.

Since n_p divides the prime q, the only possibilities are $n_p = 1$ or q. In particular, if $p \nmid q - 1$, (that is, if $q \not\equiv 1 (\text{mod } p)$), then n_p cannot equal q, so $P \trianglelefteq G$.

Let $P = \langle x\rangle$ and $Q = \langle y\rangle$. If $P \trianglelefteq G$, then since $G/C_G(P)$ is isomorphic to a subgroup of $\text{Aut}(Z_p)$ and the latter group has order $p - 1$, Lagrange's Theorem together with the observation that neither p nor q can divide $p - 1$ implies that $G = C_G(P)$. In this case $x \in P \leq Z(G)$ so x and y commute. (Alternatively, this follows immediately from Exercise 42 of Section 3.1.) This means $|xy| = pq$ (cf. the exercises in Section 2.3), hence in this case G is cyclic: $G \cong Z_{pq}$.

If $p \mid q - 1$, we shall see in Chapter 5 that there is a unique non-abelian group of order pq (in which, necessarily, $n_p = q$). We can prove the existence of this group now. Let Q be a Sylow q-subgroup of the symmetric group of degree q, S_q. By Exercise 34 in Section 3, $|N_{S_q}(Q)| = q(q - 1)$. By assumption, $p \mid q - 1$ so by Cauchy's Theorem

$N_{S_q}(Q)$ has a subgroup, P, of order p. By Corollary 15 in Section 3.2, PQ is a group of order pq. Since $C_{S_q}(Q) = Q$ (Example 2, Section 3), PQ is a non-abelian group. The essential ingredient in the uniqueness proof of PQ is Theorem 17 on the cyclicity of $\mathrm{Aut}(Z_q)$.

Example: (Groups of order 30)

Let G be a group of order 30. We show that G has a normal subgroup isomorphic to Z_{15}. We shall use this information to classify groups of order 30 in the next chapter. Note that any subgroup of order 15 is necessarily normal (since it is of index 2) and cyclic (by the preceding result) so it is only necessary to show there exists a subgroup of order 15. The quickest way of doing this is to quote Exercise 13 in Section 2. We give an alternate argument which illustrates how Sylow's Theorem can be used in conjunction with a counting of elements of prime order to produce a normal subgroup.

Let $P \in Syl_5(G)$ and let $Q \in Syl_3(G)$. If either P or Q is normal in G, by Corollary 15, Chapter 3, PQ is a group of order 15. Note also that if either P or Q is normal, then both P and Q are characteristic subgroups of PQ, and since $PQ \trianglelefteq G$, *both* P and Q are normal in G (Exercise 8(a), Section 4). Assume therefore that neither Sylow subgroup is normal. The only possibilities by Part (3) of Sylow's Theorem are $n_5 = 6$ and $n_3 = 10$. Each element of order 5 lies in a Sylow 5-subgroup, each Sylow 5-subgroup contains 4 nonidentity elements and, by Lagrange's Theorem, distinct Sylow 5-subgroups intersect in the identity. Thus the number of elements of order 5 in G is the number of nonidentity elements in one Sylow 5-subgroup times the number of Sylow 5-subgroups. This would be $4 \cdot 6 = 24$ elements of order 5. By similar reasoning, the number of elements of order 3 would be $2 \cdot 10 = 20$. This is absurd since a group of order 30 cannot contain $24 + 20 = 44$ distinct elements. One of P or Q (hence both) must be normal in G.

This sort of counting technique is frequently useful (cf. also Section 6.2) and works particularly well when the Sylow p-subgroups have order p (as in this example), since then the intersection of two distinct Sylow p-subgroups must be the identity. If the order of the Sylow p-subgroup is p^α with $\alpha \geq 2$, greater care is required in counting elements, since in this case distinct Sylow p-subgroups may have many more elements in common, i.e. the intersection may be nontrivial.

Example: (Groups of order 12)

Let G be a group of order 12. We show that either G has a normal Sylow 3-subgroup or $G \cong A_4$ (in the latter case G has a normal Sylow 2-subgroup). We shall use this information to classify groups of order 12 in the next chapter.

Suppose $n_3 \neq 1$ and let $P \in Syl_3(G)$. Since $n_3 \mid 4$ and $n_3 \equiv 1 \pmod 3$, it follows that $n_3 = 4$. Since distinct Sylow 3-subgroups intersect in the identity and each contains two elements of order 3, G contains $2 \cdot 4 = 8$ elements of order 3. Since $|G : N_G(P)| = n_3 = 4$, $N_G(P) = P$. Now G acts by conjugation on its four Sylow 3-subgroups, so this action affords a permutation representation

$$\varphi : G \to S_4$$

(note that we could also act by left multiplication on the left cosets of P and use Theorem 3). The kernel K of this action is the subgroup of G which normalizes all Sylow 3-subgroups of G. In particular, $K \leq N_G(P) = P$. Since P is not normal in G by assumption, $K = 1$, i.e. φ is injective and

$$G \cong \varphi(G) \leq S_4.$$

Since G contains 8 elements of order 3 and there are precisely 8 elements of order 3 in S_4, all contained in A_4, it follows that $\varphi(G)$ intersects A_4 in a subgroup of order at least 8. Since both groups have order 12 it follows that $\varphi(G) = A_4$, so that $G \cong A_4$.

Note that A_4 does indeed have 4 Sylow 3-subgroups (see Example 4 following Corollary 20), so that such a group G does exist. Also, let V be a Sylow 2-subgroup of A_4. Since $|V| = 4$, it contains all of the remaining elements of A_4. In particular, there cannot be another Sylow 2-subgroup. Thus $n_2(A_4) = 1$, i.e. $V \trianglelefteq A_4$ (which one can also see directly because V is the identity together with the three elements of S_4 which are products of two disjoint transpositions, that is, V is a union of conjugacy classes).

Example: (Groups of order $p^2 q$, p and q distinct primes)

Let G be a group of order $p^2 q$. We show that G has a normal Sylow subgroup (for either p or q). We shall use this information to classify some groups of this order in the next chapter (cf. Exercises 8 to 12 of Section 5.5). Let $P \in Syl_p(G)$ and let $Q \in Syl_q(G)$.

Consider first when $p > q$. Since $n_p \mid q$ and $n_p = 1 + kp$, we must have $n_p = 1$. Thus $P \trianglelefteq G$.

Consider now the case $p < q$. If $n_q = 1$, Q is normal in G. Assume therefore that $n_q > 1$, i.e. $n_q = 1 + tq$, for some $t > 0$. Now n_q divides p^2 so $n_q = p$ or p^2. Since $q > p$ we cannot have $n_q = p$, hence $n_q = p^2$. Thus

$$tq = p^2 - 1 = (p-1)(p+1).$$

Since q is prime, either $q \mid p-1$ or $q \mid p+1$. The former is impossible since $q > p$ so the latter holds. Since $q > p$ but $q \mid p+1$, we must have $q = p+1$. This forces $p = 2$, $q = 3$ and $|G| = 12$. The result now follows from the preceding example.

Groups of Order 60

We illustrate how Sylow's Theorems can be used to unravel the structure of groups of a given order even if some groups of that order may be simple. Note the technique of changing from one prime to another and the inductive process where we use results on groups of order < 60 to study groups of order 60.

Proposition 21. If $|G| = 60$ and G has more than one Sylow 5-subgroup, then G is simple.

Proof: Suppose by way of contradiction that $|G| = 60$ and $n_5 > 1$ but that there exists H a normal subgroup of G with $H \neq 1$ or G. By Sylow's Theorem the only possibility for n_5 is 6. Let $P \in Syl_5(G)$, so that $|N_G(P)| = 10$ since its index is n_5.

If $5 \mid |H|$ then H contains a Sylow 5-subgroup of G and since H is normal, it contains all 6 conjugates of this subgroup. In particular, $|H| \geq 1 + 6 \cdot 4 = 25$, and the only possibility is $|H| = 30$. This leads to a contradiction since a previous example proved that any group of order 30 has a normal (hence unique) Sylow 5-subgroup. This argument shows 5 does not divide $|H|$ for any proper normal subgroup H of G.

If $|H| = 6$ or 12, H has a normal, hence characteristic, Sylow subgroup, which is therefore also normal in G. Replacing H by this subgroup if necessary, we may assume $|H| = 2$, 3 or 4. Let $\overline{G} = G/H$, so $|\overline{G}| = 30$, 20 or 15. In each case, $\overline{G}$ has a normal subgroup $\overline{P}$ of order 5 by previous results. If we let H_1 be the complete preimage of $\overline{P}$ in G, then $H_1 \trianglelefteq G$, $H_1 \neq G$ and $5 \mid |H_1|$. This contradicts the preceding paragraph and so completes the proof.

Corollary 22. A_5 is simple.

Proof: The subgroups $\langle(1\,2\,3\,4\,5)\rangle$ and $\langle(1\,3\,2\,4\,5)\rangle$ are distinct Sylow 5-subgroups of A_5 so the result follows immediately from the proposition.

The next proposition shows that there is a unique simple group of order 60.

Proposition 23. If G is a simple group of order 60, then $G \cong A_5$.

Proof: Let G be a simple group of order 60, so $n_2 = 3, 5$ or 15. Let $P \in Syl_2(G)$ and let $N = N_G(P)$, so $|G : N| = n_2$.

First observe that G has no proper subgroup H of index less than 5, as follows: if H were a subgroup of G of index 4, 3 or 2, then, by Theorem 3, G would have a normal subgroup K contained in H with G/K isomorphic to a subgroup of S_4, S_3 or S_2. Since $K \neq G$, simplicity forces $K = 1$. This is impossible since 60 $(= |G|)$ does not divide 4! This argument shows, in particular, that $n_2 \neq 3$.

If $n_2 = 5$, then N has index 5 in G so the action of G by left multiplication on the set of left cosets of N gives a permutation representation of G into S_5. Since (as above) the kernel of this representation is a proper normal subgroup and G is simple, the kernel is 1 and G is isomorphic to a subgroup of S_5. Identify G with this isomorphic copy so that we may assume $G \leq S_5$. If G is not contained in A_5, then $S_5 = GA_5$ and, by the Second Isomorphism Theorem, $A_5 \cap G$ is of index 2 in G. Since G has no (normal) subgroup of index 2, this is a contradiction. This argument proves $G \leq A_5$. Since $|G| = |A_5|$, the isomorphic copy of G in S_5 coincides with A_5, as desired.

Finally, assume $n_2 = 15$. If for every pair of distinct Sylow 2-subgroups P and Q of G, $P \cap Q = 1$, then the number of nonidentity elements in Sylow 2-subgroups of G would be $(4-1)\cdot 15 = 45$. But $n_5 = 6$ so the number of elements of order 5 in G is $(5-1)\cdot 6 = 30$, accounting for 75 elements. This contradiction proves that there exist distinct Sylow 2-subgroups P and Q with $|P \cap Q| = 2$. Let $M = N_G(P \cap Q)$. Since P and Q are abelian (being groups of order 4), P and Q are subgroups of M and since G is simple, $M \neq G$. Thus 4 divides $|M|$ and $|M| > 4$ (otherwise, $P = M = Q$). The only possibility is $|M| = 12$, i.e. M has index 5 in G (recall M cannot have index 3 or 1). But now the argument of the preceding paragraph applied to M in place of N gives $G \cong A_5$. This leads to a contradiction in this case because $n_2(A_5) = 5$ (cf. the exercises). The proof is complete.

EXERCISES

Let G be a finite group and let p be a prime.

1. Prove that if $P \in Syl_p(G)$ and H is a subgroup of G containing P then $P \in Syl_p(H)$. Give an example to show that, in general, a Sylow p-subgroup of a subgroup of G need not be a Sylow p-subgroup of G.
2. Prove that if H is a subgroup of G and $Q \in Syl_p(H)$ then $gQg^{-1} \in Syl_p(gHg^{-1})$ for all $g \in G$.
3. Use Sylow's Theorem to prove Cauchy's Theorem. (Note that we only used Cauchy's Theorem for abelian groups — Proposition 3.21 — in the proof of Sylow's Theorem so this line of reasoning is not circular.)

4. Exhibit all Sylow 2-subgroups and Sylow 3-subgroups of D_{12} and $S_3 \times S_3$.
5. Show that a Sylow p-subgroup of D_{2n} is cyclic and normal for every odd prime p.
6. Exhibit all Sylow 3-subgroups of A_4 and all Sylow 3-subgroups of S_4.
7. Exhibit all Sylow 2-subgroups of S_4 and find elements of S_4 which conjugate one of these into each of the others.
8. Exhibit two distinct Sylow 2-subgroups of S_5 and an element of S_5 that conjugates one into the other.
9. Exhibit all Sylow 3-subgroups of $SL_2(\mathbb{F}_3)$ (cf. Exercise 9, Section 2.1).
10. Prove that the subgroup of $SL_2(\mathbb{F}_3)$ generated by $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ is the unique Sylow 2-subgroup of $SL_2(\mathbb{F}_3)$ (cf. Exercise 10, Section 2.4).
11. Show that the center of $SL_2(\mathbb{F}_3)$ is the group of order 2 consisting of $\pm I$, where I is the identity matrix. Prove that $SL_2(\mathbb{F}_3)/Z(SL_2(\mathbb{F}_3)) \cong A_4$. [Use facts about groups of order 12.]
12. Let $2n = 2^a k$ where k is odd. Prove that the number of Sylow 2-subgroups of D_{2n} is k. [Prove that if $P \in Syl_2(D_{2n})$ then $N_{D_{2n}}(P) = P$.]
13. Prove that a group of order 56 has a normal Sylow p-subgroup for some prime p dividing its order.
14. Prove that a group of order 312 has a normal Sylow p-subgroup for some prime p dividing its order.
15. Prove that a group of order 351 has a normal Sylow p-subgroup for some prime p dividing its order.
16. Let $|G| = pqr$, where p, q and r are primes with $p < q < r$. Prove that G has a normal Sylow subgroup for either p, q or r.
17. Prove that if $|G| = 105$ then G has a normal Sylow 5-subgroup and a normal Sylow 7-subgroup.
18. Prove that a group of order 200 has a normal Sylow 5-subgroup.
19. Prove that if $|G| = 6545$ then G is not simple.
20. Prove that if $|G| = 1365$ then G is not simple.
21. Prove that if $|G| = 2907$ then G is not simple.
22. Prove that if $|G| = 132$ then G is not simple.
23. Prove that if $|G| = 462$ then G is not simple.
24. Prove that if G is a group of order 231 then $Z(G)$ contains a Sylow 11-subgroup of G and a Sylow 7-subgroup is normal in G.
25. Prove that if G is a group of order 385 then $Z(G)$ contains a Sylow 7-subgroup of G and a Sylow 11-subgroup is normal in G.
26. Let G be a group of order 105. Prove that if a Sylow 3-subgroup of G is normal then G is abelian.
27. Let G be a group of order 315 which has a normal Sylow 3-subgroup. Prove that $Z(G)$ contains a Sylow 3-subgroup of G and deduce that G is abelian.
28. Let G be a group of order 1575. Prove that if a Sylow 3-subgroup of G is normal then a Sylow 5-subgroup and a Sylow 7-subgroup are normal. In this situation prove that G is abelian.
29. If G is a non-abelian simple group of order < 100, prove that $G \cong A_5$. [Eliminate all orders but 60.]

30. How many elements of order 7 must there be in a simple group of order 168?

31. For $p = 2, 3$ and 5 find $n_p(A_5)$ and $n_p(S_5)$. [Note that $A_4 \leq A_5$.]

32. Let P be a Sylow p-subgroup of H and let H be a subgroup of K. If $P \trianglelefteq H$ and $H \trianglelefteq K$, prove that P is normal in K. Deduce that if $P \in Syl_p(G)$ and $H = N_G(P)$, then $N_G(H) = H$ (in words: *normalizers of Sylow p-subgroups are self-normalizing*).

33. Let P be a normal Sylow p-subgroup of G and let H be any subgroup of G. Prove that $P \cap H$ is the unique Sylow p-subgroup of H.

34. Let $P \in Syl_p(G)$ and assume $N \trianglelefteq G$. Use the conjugacy part of Sylow's Theorem to prove that $P \cap N$ is a Sylow p-subgroup of N. Deduce that PN/N is a Sylow p-subgroup of G/N (note that this may also be done by the Second Isomorphism Theorem — cf. Exercise 9, Section 3.3).

35. Let $P \in Syl_p(G)$ and let $H \leq G$. Prove that $gPg^{-1} \cap H$ is a Sylow p-subgroup of H for some $g \in G$. Give an explicit example showing that $hPh^{-1} \cap H$ is not necessarily a Sylow p-subgroup of H for any $h \in H$ (in particular, we cannot always take $g = 1$ in the first part of this problem, as we could when H was normal in G).

36. Prove that if N is a normal subgroup of G then $n_p(G/N) \leq n_p(G)$.

37. Let R be a normal p-subgroup of G (not necessarily a Sylow subgroup).
    (a) Prove that R is contained in every Sylow p-subgroup of G.
    (b) If S is another normal p-subgroup of G, prove that RS is also a normal p-subgroup of G.
    (c) The subgroup $O_p(G)$ is defined to be the group generated by all normal p-subgroups of G. Prove that $O_p(G)$ is the unique largest normal p-subgroup of G and $O_p(G)$ equals the intersection of all Sylow p-subgroups of G.
    (d) Let $\overline{G} = G/O_p(G)$. Prove that $O_p(\overline{G}) = \overline{1}$ (i.e. $\overline{G}$ has no nontrivial normal p-subgroup).

38. Use the method of proof in Sylow's Theorem to show that if n_p is not congruent to $1 (\text{mod } p^2)$ then there are distinct Sylow p-subgroups P and Q of G such that $|P : P \cap Q| = |Q : P \cap Q| = p$.

39. Show that the subgroup of strictly upper triangular matrices in $GL_n(\mathbb{F}_p)$ (cf. Exercise 17, Section 2.1) is a Sylow p-subgroup of this finite group. [Use the order formula in Section 1.4 to find the order of a Sylow p-subgroup of $GL_n(\mathbb{F}_p)$.]

40. Prove that the number of Sylow p-subgroups of $GL_2(\mathbb{F}_p)$ is $p + 1$. [Exhibit two distinct Sylow p-subgroups.]

41. Prove that $SL_2(\mathbb{F}_4) \cong A_5$ (cf. Exercise 9, Section 2.1 for the definition of $SL_2(\mathbb{F}_4)$).

42. Prove that the group of rigid motions in $\mathbb{R}^3$ of an icosahedron is isomorphic to A_5. [Recall that the order of this group is 60: Exercise 13, Section 1.2.]

43. Prove that the group of rigid motions in $\mathbb{R}^3$ of a dodecahedron is isomorphic to A_5. (As with the cube and the tetrahedron, the icosahedron and the dodecahedron are dual solids.) [Recall that the order of this group is 60: Exercise 12, Section 1.2.]

44. Let p be the smallest prime dividing the order of the finite group G. Prove that if $P \in Syl_p(G)$ and P is cyclic then $N_G(P) = C_G(P)$.

45. Find generators for a Sylow p-subgroup of S_{2p}, where p is an odd prime. Show that this is an abelian group of order p^2.

46. Find generators for a Sylow p-subgroup of S_{p^2}, where p is a prime. Show that this is a non-abelian group of order p^{p+1}.

47. Write and execute a computer program which
    (i) gives each odd number $n < 10,000$ that is not a power of a prime and that has some

prime divisor p such that n_p is not forced to be 1 for all groups of order n by the congruence condition of Sylow's Theorem, and

(ii) gives for each n in (i) the factorization of n into prime powers and gives the list of all permissible values of n_p for all primes p dividing n (i.e. those values not ruled out by Part 3 of Sylow's Theorem).

48. Carry out the same process as in the preceding exercise for all even numbers less than 1,000. Explain the relative lengths of the lists versus the number of integers tested.

49. Prove that if $|G| = 2^n m$ where m is odd and G has a cyclic Sylow 2-subgroup then G has a normal subgroup of order m. [Use induction and Exercises 11 and 12 in Section 2.]

50. Let $P \in Syl_p(G)$. Prove that two elements of $Z(P)$ are conjugate in G if and only if they are conjugate in $N_G(P)$.

51. Let P be a Sylow p-subgroup of G and let M be any subgroup of G which contains $N_G(P)$. Prove that $|G : M| \equiv 1 (\text{mod } p)$.

The following sequence of exercises leads to the classification of all numbers n with the property that every group of order n is cyclic (for example, $n = 15$ is such an integer). These arguments are a vastly simplified prototype for the proof that every group of odd order is solvable in the sense that they use the *structure* (commutativity) of the proper subgroups and their *embedding* in the whole group (we shall see that distinct maximal subgroups intersect in the identity) to obtain a contradiction by counting arguments. In the proof that groups of odd order are solvable one uses induction to reduce to the situation in which a minimal counterexample is a simple group—but here every proper subgroup is solvable (not abelian as in our situation). The analysis of the structure and embedding of the maximal subgroups in this situation is much more complicated and the counting arguments are (roughly speaking) replaced by character theory arguments (as will be discussed in Part V).

52. Suppose G is a finite simple group in which every proper subgroup is abelian. Prove that if M and N are distinct maximal subgroups of G then $M \cap N = 1$. [See Exercise 23 in Section 3.]

53. Use the preceding exercise to prove that if G is any non-abelian group in which every proper subgroup is abelian then G is not simple. [Let G be a counterexample to this assertion and use Exercise 24 in Section 3 to show that G has more than one conjugacy class of maximal subgroups. Use the method of Exercise 23 in Section 3 to count the elements which lie in all conjugates of M and N, where M and N are nonconjugate maximal subgroups of G; show that this gives more than $|G|$ elements.]

54. Prove the following classification: if G is a finite group of order $p_1 p_2 \ldots p_r$ where the p_i's are distinct primes such that p_i does not divide $p_j - 1$ for all i and j, then G is cyclic. [By induction, every proper subgroup of G is cyclic, so G is not simple by the preceding exercise. If N is a nontrivial proper normal subgroup, N is cyclic and G/N acts as automorphisms of N. Use Proposition 16 to show that $N \leq Z(G)$ and use induction to show $G/Z(G)$ is cyclic, hence G is abelian by Exercise 36 of Section 3.1.]

55. Prove the converse to the preceding exercise: if $n \geq 2$ is an integer such that every group of order n is cyclic, then $n = p_1 p_2 \ldots p_r$ is a product of distinct primes and p_i does not divide $p_j - 1$ for all i, j. [If n is not of this form, construct noncyclic groups of order n using direct products of noncyclic groups of order p^2 and pq, where $p \mid q - 1$.]

56. If G is a finite group in which every proper subgroup is abelian, show that G is solvable.

4.6 THE SIMPLICITY OF A_n

There are a number of proofs of the simplicity of A_n, $n \geq 5$. The most elementary involves showing A_n is generated by 3-cycles. Then one shows that a normal subgroup must contain one 3-cycle hence must contain all the 3-cycles so cannot be a proper subgroup. We include a less computational approach.

Note that A_3 is an abelian simple group and that A_4 is not simple ($n_2(A_4) = 1$).

Theorem 24. A_n is simple for all $n \geq 5$.

Proof: By induction on n. The result has already been established for $n = 5$, so assume $n \geq 6$ and let $G = A_n$. Assume there exists $H \trianglelefteq G$ with $H \neq 1$ or G. For each $i \in \{1, 2, \ldots, n\}$ let G_i be the stabilizer of i in the natural action of G on $i \in \{1, 2, \ldots, n\}$. Thus $G_i \leq G$ and $G_i \cong A_{n-1}$. By induction, G_i is simple for $1 \leq i \leq n$.

Suppose first that there is some $\tau \in H$ with $\tau \neq 1$ but $\tau(i) = i$ for some $i \in \{1, 2, \ldots, n\}$. Since $\tau \in H \cap G_i$ and $H \cap G_i \trianglelefteq G_i$, by the simplicity of G_i we must have $H \cap G_i = G_i$, that is

$$G_i \leq H.$$

By Exercise 2 of Section 1, $\sigma G_i \sigma^{-1} = G_{\sigma(i)}$, so for all i, $\sigma G_i \sigma^{-1} \leq \sigma H \sigma^{-1} = H$. Thus

$$G_j \leq H, \quad \text{for all } j \in \{1, 2, \ldots, n\}.$$

Any $\lambda \in A_n$ may be written as a product of an even number, $2t$, of transpositions, so

$$\lambda = \lambda_1 \lambda_2 \ldots \lambda_t,$$

where λ_k is a product of two transpositions. Since $n > 4$ each $\lambda_k \in G_j$, for some j, hence

$$G = \langle G_1, G_2, \ldots, G_n \rangle \leq H,$$

which is a contradiction. Therefore if $\tau \neq 1$ is an element of H then $\tau(i) \neq i$ for all $i \in \{1, 2, \ldots, n\}$, i.e. no nonidentity element of H fixes any element of $\{1, 2, \ldots, n\}$.

It follows that if τ_1, τ_2 are elements of H with

$$\tau_1(i) = \tau_2(i) \text{ for some } i, \text{ then } \tau_1 = \tau_2 \qquad (*)$$

since then $\tau_2^{-1}\tau_1(i) = i$.

Suppose there exists a $\tau \in H$ such that the cycle decomposition of τ contains a cycle of length ≥ 3, say

$$\tau = (a_1\, a_2\, a_3 \ldots)(b_1\, b_2\, \ldots) \ldots .$$

Let $\sigma \in G$ be an element with $\sigma(a_1) = a_1$, $\sigma(a_2) = a_2$ but $\sigma(a_3) \neq a_3$ (note that such a σ exists in A_n since $n \geq 5$. By Proposition 10

$$\tau_1 = \sigma\tau\sigma^{-1} = (a_1\, a_2\, \sigma(a_3) \ldots)(\sigma(b_1)\, \sigma(b_2) \ldots)$$

so τ and τ_1 are distinct elements of H with $\tau(a_1) = \tau_1(a_1) = a_2$, contrary to $(*)$. This proves that only 2-cycles can appear in the cycle decomposition of nonidentity elements of H.

Let $\tau \in H$ with $\tau \neq 1$, so that

$$\tau = (a_1\, a_2)(a_3\, a_4)(a_5\, a_6)\ldots$$

(note that $n \geq 6$ is used here). Let $\sigma = (a_1\, a_2)(a_3\, a_5) \in G$. Then

$$\tau_1 = \sigma\tau\sigma^{-1} = (a_1\, a_2)(a_5\, a_4)(a_3\, a_6)\ldots$$

hence τ and τ_1 are distinct elements of H with $\tau(a_1) = \tau_1(a_1) = a_2$, again contrary to $(*)$. This completes the proof of the simplicity of A_n.

EXERCISES

Let G be a group and let Ω be an infinite set.

1. Prove that A_n does not have a proper subgroup of index $< n$ for all $n \geq 5$.
2. Find all normal subgroups of S_n for all $n \geq 5$.
3. Prove that A_n is the only proper subgroup of index $< n$ in S_n for all $n \geq 5$.
4. Prove that A_n is generated by the set of all 3-cycles for each $n \geq 3$.
5. Prove that if there exists a chain of subgroups: $G_1 \leq G_2 \leq \ldots \leq G$ such that $G = \cup_{i=1}^{\infty} G_i$ and each G_i is simple then G is simple.
6. Let D be the subgroup of S_Ω consisting of permutations which move only a finite number of elements of Ω (described in Exercise 17 in Section 3) and let A be the set of all elements $\sigma \in D$ such that σ acts as an even permutation on the (finite) set of points it moves. Prove that A is an infinite simple group. [Use the preceding exercise.]
7. Under the notation of the preceding exercise prove that if $H \trianglelefteq S_\Omega$ and $H \neq 1$ then $A \leq H$, i.e. A is the unique (nontrivial) minimal normal subgroup of S_Ω.
8. Under the notation of the preceding two exercises prove that $|D| = |A| = |\Omega|$. Deduce that

$$\text{if } S_\Omega \cong S_\Delta \text{ then } |\Omega| = |\Delta|.$$

[Use the fact that D is generated by transpositions. You may assume that countable unions and finite direct products of sets of cardinality $|\Omega|$ also have cardinality $|\Omega|$.]

5

Direct and Semidirect Products and Abelian Groups

In this chapter we consider two of the easier methods for constructing larger groups from smaller ones, namely the notions of direct and semidirect products. This allows us to state the Fundamental Theorem on Finitely Generated Abelian Groups, which in particular completely classifies all finite abelian groups.

5.1 DIRECT PRODUCTS

We begin with the definition of the direct product of a finite and of a countable number of groups (the direct product of an arbitary collection of groups is considered in the exercises).

Definition

(1) The *direct product* $G_1 \times G_2 \times \cdots \times G_n$ of the groups $(G_1, \star_1), (G_2, \star_2), \ldots, (G_n, \star_n)$ is the set of n-tuples $(g_1, g_2, \ldots, g_n)$ where $g_i \in G_i$ with operation defined componentwise:

$$(g_1,\ g_2, \ldots, g_n) \star (h_1,\ h_2, \ldots, h_n) = (g_1 \star_1 h_1,\ g_2 \star_2 h_2, \ldots, g_n \star_n h_n).$$

(2) Similarly, the *direct product* $G_1 \times G_2 \times \cdots$ of the groups $(G_1, \star_1), (G_2, \star_2), \ldots$ is the set of sequences $(g_1, g_2, \ldots)$ where $g_i \in G_i$ with operation defined componentwise:

$$(g_1,\ g_2, \ldots) \star (h_1,\ h_2, \ldots) = (g_1 \star_1 h_1,\ g_2 \star_2 h_2, \ldots).$$

Although the operations may be different in each of the factors of a direct product, we shall, as usual, write all abstract groups multiplicatively, so that the operation in (1) above, for example, becomes simply

$$(g_1,\ g_2, \ldots, g_n)(h_1,\ h_2, \ldots, h_n) = (g_1h_1,\ g_2h_2, \ldots, g_nh_n).$$

Examples

(1) Suppose $G_i = \mathbb{R}$ (operation addition) for $i = 1, 2, \ldots, n$. Then $\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ (n-factors) is the familiar Euclidean n-space $\mathbb{R}^n$ with usual vector addition:

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n).$$

(2) To illustrate that groups forming the direct product (and corresponding operations) may be completely general, let $G_1 = \mathbb{Z}$, let $G_2 = S_3$ and let $G_3 = GL_2(\mathbb{R})$, where the group operations are addition, composition, and matrix multiplication, respectively. Then the operation in $G_1 \times G_2 \times G_3$ is defined by

$$(n, \sigma, \begin{pmatrix} a & b \\ c & d \end{pmatrix})\,(m, \tau, \begin{pmatrix} p & q \\ r & s \end{pmatrix}) = (n+m, \sigma \circ \tau, \begin{pmatrix} ap+br & aq+bs \\ cp+dr & cq+ds \end{pmatrix}).$$

Proposition 1. If $G_1, \ldots, G_n$ are groups, their direct product is a group of order $|G_1|\,|G_2| \ldots |G_n|$ (if any G_i is infinite, so is the direct product).

Proof: Let $G = G_1 \times G_2 \times \cdots \times G_n$. The proof that the group axioms hold for G is straightforward since each axiom is a consequence of the fact that the same axiom holds in each factor, G_i, and the operation on G is defined componentwise. For example, the associative law is verified as follows:

Let $(a_1, a_2, \ldots, a_n)$, $(b_1, b_2, \ldots, b_n)$, and $(c_1, c_2, \ldots, c_n) \in G$. Then

$$\begin{aligned}(a_1, a_2, \ldots, a_n)\big[(b_1, b_2, \ldots, b_n)(c_1, c_2, \ldots, c_n)\big] &= (a_1, a_2, \ldots, a_n)(b_1c_1, b_2c_2, \ldots, b_nc_n) \\ &= (a_1(b_1c_1), a_2(b_2c_2), \ldots, a_n(b_nc_n)) \\ &= ((a_1b_1)c_1, (a_2b_2)c_2, \ldots, (a_nb_n)c_n) \\ &= \big[(a_1, a_2, \cdots, a_n)(b_1, b_2, \ldots, b_n)\big](c_1, c_2, \ldots, c_n\end{aligned}$$

where in the third step we have used the associative law in each component. The remaining verification that the direct product is a group is similar: the identity of G is the n-tuple $(1_1, 1_2, \ldots, 1_n)$, where 1_i is the identity of G_i and the inverse of $(g_1, g_2, \ldots, g_n)$ is $(g_1^{-1}, g_2^{-1}, \ldots, g_n^{-1})$, where g_i^{-1} is the inverse of g_i in G_i.

The formula for the order of G is clear.

If the factors of the direct product are rearranged, the resulting direct product is isomorphic to the original one (cf. Exercise 7).

The next proposition shows that a direct product, $G_1 \times G_2 \times \cdots \times G_n$, contains an isomorphic copy of each G_i. One can think of these specific copies as the "coordinate axes" of the direct product since, in the case of $\mathbb{R} \times \mathbb{R}$, they coincide with the x and y axes. One should be careful, however, not to think of these "coordinate axes" as the *only* copies of the groups G_i in the direct product. For example in $\mathbb{R} \times \mathbb{R}$ any line through the origin is a subgroup of $\mathbb{R} \times \mathbb{R}$ isomorphic to $\mathbb{R}$ (and $\mathbb{R} \times \mathbb{R}$ has infinitely many pairs of lines which are coordinate axes, viz. any rotation of a given coordinate system). The second part of the proposition shows that there are *projection homomorphisms* onto each of the components.

Proposition 2. Let $G_1, G_2, \ldots, G_n$ be groups and let $G = G_1 \times \cdots \times G_n$ be their direct product.

(1) For each fixed i the set of elements of G which have the identity of G_j in the jth position for all $j \neq i$ and arbitrary elements of G_i in position i is a subgroup of G isomorphic to G_i :

$$G_i \cong \{(1, 1, \ldots, 1, g_i, 1, \ldots, 1) \mid g_i \in G_i\}.$$

(here g_i appears in the ith position). If we identify G_i with this subgroup, then $G_i \trianglelefteq G$ and

$$G/G_i \cong G_1 \times \cdots \times G_{i-1} \times G_{i+1} \times \cdots \times G_n.$$

(2) For each fixed i define $\pi_i : G \to G_i$ by

$$\pi_i((g_1, g_2, \ldots, g_n)) = g_i.$$

Then π_i is a surjective homomorphism with

$$\begin{aligned} \ker \pi_i &= \{(g_1, \ldots, g_{i-1}, 1, g_{i+1}, \ldots, g_n) \mid g_j \in G_j \text{ for all } j \neq i\} \\ &\cong G_1 \times \cdots \times G_{i-1} \times G_{i+1} \times \cdots \times G_n \end{aligned}$$

(here the 1 appears in position i).

(3) Under the identifications in part (1), if $x \in G_i$ and $y \in G_j$ for some $i \neq j$, then $xy = yx$.

Proof: (1) Since the operation in G is defined componentwise, it follows easily from the subgroup criterion that $\{(1, 1, \ldots, 1, g_i, 1, \ldots, 1) \mid g_i \in G_i\}$ is a subgroup of G. Furthermore, the map $g_i \mapsto (1, 1, \ldots, 1, g_i, 1, \ldots, 1)$ is seen to be an isomorphism of G_i with this subgroup. Identify G_i with this isomorphic copy in G.

To prove the remaining parts of (1) consider the map $\varphi : G \to G_1 \times \cdots \times G_{i-1} \times G_{i+1} \times \cdots \times G_n$ defined by

$$\varphi(g_1, g_2, \ldots, g_n) = (g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n)$$

(i.e. φ erases the ith component of G). The map φ is a homomorphism since

$$\begin{aligned} \varphi((g_1, \ldots, g_n)(h_1, \ldots, h_n)) &= \varphi((g_1h_1, \ldots, g_nh_n)) \\ &= (g_1h_1, \ldots, g_{i-1}h_{i-1}, g_{i+1}h_{i+1}, \ldots, g_nh_n) \\ &= (g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_n)(h_1, \ldots, h_{i-1}, h_{i+1}, \ldots, h_n) \\ &= \varphi((g_1, \ldots, g_n))\varphi((h_1, \ldots, h_n)). \end{aligned}$$

Since the entries in position j are arbitrary elements of G_j for all j, φ is surjective. Furthermore,

$$\ker \varphi = \{(g_1, \ldots, g_n) \mid g_j = 1 \text{ for all } j \neq i\} = G_i.$$

This proves that G_i is a normal subgroup of G (in particular, it again proves this copy of G_i is a subgroup) and the First Isomorphism Theorem gives the final assertion of part (1).

In (2) the argument that π_i is a surjective homomorphism and the kernel is the subgroup described is very similar to that in part (1), so the details are left to the reader.

In part (3) if $x = (1, \ldots, 1, g_i, 1, \ldots, 1)$ and $y = (1, \ldots, 1, g_j, 1, \ldots, 1)$, where the indicated entries appear in positions i, j respectively, then

$$xy = (1, \ldots, 1, g_i, 1, \ldots, 1, g_j, 1, \ldots, 1) = yx$$

(where the notation is chosen so that $i < j$). This completes the proof.

A generalization of this proposition appears as Exercise 2.

We shall continue to identify the "coordinate axis" subgroups described in part (1) of the proposition with their isomorphic copies, the G_i's. The ith such subgroup is often called the ith *component* or i^{th} *factor* of G. For instance, when we wish to calculate in $Z_n \times Z_m$ we can let x be a generator of the first factor, let y be a generator of the second factor and write the elements of $Z_n \times Z_m$ in the form $x^a y^b$. This replaces the formal ordered pairs $(x, 1)$ and $(1, y)$ with x and y (so $x^a y^b$ replaces (x^a, y^b)).

Examples

(1) Under the notation of Proposition 2 it follows from part (3) that if $x_i \in G_i$, $1 \leq i \leq n$, then for all $k \in \mathbb{Z}$

$$(x_1 x_2 \ldots x_n)^k = x_1^k x_2^k \ldots x_n^k.$$

Since the order of $x_1 x_2 \ldots x_n$ is the smallest positive integer k such that $x_i^k = 1$ for all i, we see that

$$|x_1 x_2 \ldots x_n| = \text{l.c.m.}(|x_1|, |x_2|, \ldots, |x_n|)$$

(where this order is infinite if and only if one of the x_i's has infinite order).

(2) Let p be a prime and for $n \in \mathbb{Z}^+$ consider

$$E_{p^n} = Z_p \times Z_p \times \cdots \times Z_p \qquad (n \text{ factors}).$$

Then E_{p^n} is an abelian group of order p^n with the property that $x^p = 1$ for all $x \in E_{p^n}$. This group is the *elementary abelian* group of order p^n described in Section 4.4.

(3) For p a prime, we show that the elementary abelian group of order p^2 has exactly $p+1$ subgroups of order p (in particular, there are more than the two obvious ones). Let $E = E_{p^2}$. Since each nonidentity element of E has order p, each of these generates a cyclic subgroup of E of order p. By Lagrange's Theorem distinct subgroups of order p intersect trivially. Thus the $p^2 - 1$ nonidentity elements of E are partitioned into subsets of size $p - 1$ (i.e. each of these subsets consists of the nonidentity elements of some subgroup of order p). There must therefore be

$$\frac{p^2 - 1}{p - 1} = p + 1$$

subgroups of order p. When $p = 2$, E is the Klein 4-group which we have already seen has 3 subgroups of order 2 (cf. also Exercises 10 and 11).

EXERCISES

1. Show that the center of a direct product is the direct product of the centers:

$$Z(G_1 \times G_2 \times \cdots \times G_n) = Z(G_1) \times Z(G_2) \times \cdots \times Z(G_n).$$

Deduce that a direct product of groups is abelian if and only if each of the factors is abelian.

2. Let $G_1, G_2, \ldots, G_n$ be groups and let $G = G_1 \times \cdots \times G_n$. Let I be a proper, nonempty subset of $\{1, \ldots, n\}$ and let $J = \{1, \ldots, n\} - I$. Define G_I to be the set of elements of G that have the identity of G_j in position j for all $j \in J$.

(a) Prove that G_I is isomorphic to the direct product of the groups G_i, $i \in I$.

(b) Prove that G_I is a normal subgroup of G and $G/G_I \cong G_J$.
(c) Prove that $G \cong G_I \times G_J$.

3. Under the notation of the preceding exercise let I and K be any disjoint nonempty subsets of $\{1, 2, \ldots, n\}$ and let G_I and G_K be the subgroups of G defined above. Prove that $xy = yx$ for all $x \in G_I$ and all $y \in G_K$.

4. Let A and B be finite groups and let p be a prime. Prove that any Sylow p-subgroup of $A \times B$ is of the form $P \times Q$, where $P \in Syl_p(A)$ and $Q \in Syl_p(B)$. Prove that $n_p(A \times B) = n_p(A)n_p(B)$. Generalize both of these results to a direct product of any finite number of finite groups (so that the number of Sylow p-subgroups of a direct product is the product of the numbers of Sylow p-subgroups of the factors).

5. Exhibit a nonnormal subgroup of $Q_8 \times Z_4$ (note that every subgroup of each factor is normal).

6. Show that all subgroups of $Q_8 \times E_{2^n}$ are normal.

7. Let $G_1, G_2, \ldots, G_n$ be groups and let π be a fixed element of S_n. Prove that the map

$$\varphi_\pi : G_1 \times G_2 \times \cdots \times G_n \to G_{\pi^{-1}(1)} \times G_{\pi^{-1}(2)} \times \cdots \times G_{\pi^{-1}(n)}$$

defined by

$$\varphi_\pi(g_1, g_2, \ldots, g_n) = (g_{\pi^{-1}(1)}, g_{\pi^{-1}(2)}, \ldots, g_{\pi^{-1}(n)})$$

is an isomorphism (so that changing the order of the factors in a direct product does not change the isomorphism type).

8. Let $G_1 = G_2 = \cdots = G_n$ and let $G = G_1 \times \cdots \times G_n$. Under the notation of the preceding exercise show that $\varphi_\pi \in \mathrm{Aut}(G)$. Show also that the map $\pi \mapsto \varphi_\pi$ is an injective homomorphism of S_n into $\mathrm{Aut}(G)$. (In particular, $\varphi_{\pi_1} \circ \varphi_{\pi_2} = \varphi_{\pi_1\pi_2}$. It is at this point that the π^{-1}'s in the definition of φ_π are needed. The underlying reason for this is because if e_i is the n-tuple with 1 in position i and zeros elsewhere, $1 \le i \le n$, then S_n acts on $\{e_1, \ldots, e_n\}$ by $\pi \cdot e_i = e_{\pi(i)}$; this is a left group action. If the n-tuple $(g_1, \ldots, g_n)$ is represented by $g_1e_1 + \cdots + g_ne_n$, then this left group action on $\{e_1, \ldots, e_n\}$ extends to a left group action on sums by

$$\pi \cdot (g_1e_1 + g_2e_2 + \cdots + g_ne_n) = g_1e_{\pi(1)} + g_2e_{\pi(2)} + \cdots + g_ne_{\pi(n)}.$$

The coefficient of $e_{\pi(i)}$ on the right hand side is g_i, so the coefficient of e_i is $g_{\pi^{-1}(i)}$. Thus the right hand side may be rewritten as $g_{\pi^{-1}(1)}e_1 + g_{\pi^{-1}(2)}e_2 + \cdots + g_{\pi^{-1}(n)}e_n$, which is precisely the sum attached to the n-tuple $(g_{\pi^{-1}(1)}, g_{\pi^{-1}(2)}, \ldots, g_{\pi^{-1}(n)})$. In other words, any permutation of the " position vectors" $e_1, \ldots, e_n$ (which fixes their coefficients) is the same as the inverse permutation on the coefficients (fixing the e_i's). If one uses π's in place of π^{-1}'s in the definition of φ_π then the map $\pi \mapsto \varphi_\pi$ is not necessarily a homomorphism — it corresponds to a *right* group action.)

9. Let G_i be a field F for all i and use the preceding exercise to show that the set of $n \times n$ matrices with one 1 in each row and each column is a subgroup of $GL_n(F)$ isomorphic to S_n (these matrices are called *permutation matrices* since they simply permute the standard basis $e_1, \ldots, e_n$ (as above) of the n-dimensional vector space $F \times F \times \cdots \times F$).

10. Let p be a prime. Let A and B be two cyclic groups of order p with generators x and y respectively. Set $E = A \times B$ so that E is the elementary abelian group of order p^2: E_{p^2}. Prove that the distinct subgroups of E of order p are

$$\langle x \rangle, \quad \langle xy \rangle, \quad \langle xy^2 \rangle, \quad \ldots, \quad \langle xy^{p-1} \rangle, \quad \langle y \rangle$$

(note that there are $p + 1$ of them).

11. Let p be a prime and let $n \in \mathbb{Z}^+$. Find a formula for the number of subgroups of order p in the elementary abelian group E_{p^n}.

12. Let A and B be groups. Assume $Z(A)$ contains a subgroup Z_1 and $Z(B)$ contains a subgroup Z_2 with $Z_1 \cong Z_2$. Let this isomorphism be given by the map $x_i \mapsto y_i$ for all $x_i \in Z_1$. A *central product* of A and B is a quotient

$$(A \times B)/Z \quad \text{where} \quad Z = \{(x_i, y_i^{-1}) \mid x_i \in Z_1\}$$

and is denoted by $A * B$ – it is not unique since it depends on Z_1, Z_2 and the isomorphism between them. (Think of $A * B$ as the direct product of A and B "collapsed" by identifying each element $x_i \in Z_1$ with its corresponding element $y_i \in Z_2$.)

(a) Prove that the images of A and B in the quotient group $A * B$ are isomorphic to A and B respectively and that these images intersect in a central subgroup isomorphic to Z_1. Find $|A * B|$.

(b) Let $Z_4 = \langle x \rangle$. Let $D_8 = \langle r, s \rangle$ and $Q_8 = \langle i, j \rangle$ be given by their usual generators and relations. Let $Z_4 * D_8$ be the central product of Z_4 and D_8 which identifies x^2 and r^2 (i.e. $Z_1 = \langle x^2 \rangle$, $Z_2 = \langle r^2 \rangle$ and the isomorphism is $x^2 \mapsto r^2$) and let $Z_4 * Q_8$ be the central product of Z_4 and Q_8 which identifies x^2 and -1. Prove that $Z_4 * D_8 \cong Z_4 * Q_8$.

13. Give presentations for the groups $Z_4 * D_8$ and $Z_4 * Q_8$ constructed in the preceding exercise.

14. Let $G = A_1 \times A_2 \times \cdots \times A_n$ and for each i let B_i be a normal subgroup of A_i. Prove that $B_1 \times B_2 \times \cdots \times B_n \trianglelefteq G$ and that

$$(A_1 \times A_2 \times \cdots \times A_n)/\big(B_1 \times B_2 \times \cdots \times B_n\big) \cong (A_1/B_1) \times (A_2/B_2) \times \cdots \times (A_n/B_n).$$

The following exercise describes the direct product of an arbitrary collection of groups. The terminology for the Cartesian product of an arbitrary collection of sets may be found in the Appendix.

15. Let I be any nonempty index set and let $(G_i, \star_i)$ be a group for each $i \in I$. The *direct product* of the groups G_i, $i \in I$ is the set $G = \prod_{i \in I} G_i$ (the Cartesian product of the G_i's) with a binary operation defined as follows: if $\prod a_i$ and $\prod b_i$ are elements of G, then their product in G is given by

$$\left(\prod_{i \in I} a_i\right)\left(\prod_{i \in I} b_i\right) = \prod_{i \in I} (a_i \star_i b_i)$$

(i.e. the group operation in the direct product is defined componentwise).

(a) Show that this binary operation is well defined and associative.

(b) Show that the element $\prod 1_i$ satisfies the axiom for the identity of G, where 1_i is the identity of G_i for all i.

(c) Show that the element $\prod a_i^{-1}$ is the inverse of $\prod a_i$, where the inverse of each component element a_i is taken in the group G_i.

Conclude that the direct product is a group.

(Note that if $I = \{1, 2, \ldots, n\}$, this definition of the direct product is the same as the n-tuple definition in the text.)

16. State and prove the generalization of Proposition 2 to arbitrary direct products.

17. Let I be any nonempty index set and let G_i be a group for each $i \in I$. The *restricted direct product* or *direct sum* of the groups G_i is the set of elements of the direct product which are the identity in all but finitely many components, that is, the set of all elements $\prod a_i \in \prod_{i \in I} G_i$ such that $a_i = 1_i$ for all but a finite number of $i \in I$.

(a) Prove that the restricted direct product is a subgroup of the direct product.
(b) Prove that the restricted direct product is normal in the direct product.
(c) Let $I = \mathbb{Z}^+$ and let p_i be the ith integer prime. Show that if $G_i = \mathbb{Z}/p_i\mathbb{Z}$ for all $i \in \mathbb{Z}^+$, then every element of the restricted direct product of the G_i's has finite order but $\prod_{i \in \mathbb{Z}^+} G_i$ has elements of infinite order. Show that in this example the restricted direct product is the torsion subgroup of the direct product (cf. Exercise 6, Section 2.1).

18. In each of (a) to (c) give an example of a group with the specified properties:
(a) an infinite group in which every element has order 1 or 2
(b) an infinite group in which every element has finite order but for each positive integer n there is an element of order n
(c) a group with an element of infinite order and an element of order 2
(d) a group G such that every finite group is isomorphic to some subgroup of G
(e) a nontrivial group G such that $G \cong G \times G$.

5.2 THE FUNDAMENTAL THEOREM OF FINITELY GENERATED ABELIAN GROUPS

Definition

(1) A group G is *finitely generated* if there is a finite subset A of G such that $G = \langle A \rangle$.
(2) For each $r \in \mathbb{Z}$ with $r \geq 0$, let $\mathbb{Z}^r = \mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z}$ be the direct product of r copies of the group $\mathbb{Z}$, where $\mathbb{Z}^0 = 1$. The group $\mathbb{Z}^r$ is called the *free abelian group of rank* r.

Note that any finite group G is, a fortiori, finitely generated: simply take $A = G$ as a set of generators. Also, $\mathbb{Z}^r$ is finitely generated by $e_1, e_2, \ldots, e_r$, where e_i is the n-tuple with 1 in position i and zeros elsewhere. We can now state the fundamental classification theorem for (finitely generated) abelian groups:

Theorem 3. (The Fundamental Theorem of Finitely Generated Abelian Groups) Let G be a finitely generated abelian group. Then

(1)
$$G \cong \mathbb{Z}^r \times Z_{n_1} \times Z_{n_2} \times \cdots \times Z_{n_s},$$
for some integers $r, n_1, n_2, \ldots, n_s$ satisfying the following conditions:
(a) $r \geq 0$ and $n_j \geq 2$ for all j, and
(b) $n_{i+1} \mid n_i$ for $1 \leq i \leq s-1$

(2) the expression in (1) is unique: if $G \cong \mathbb{Z}^t \times Z_{m_1} \times Z_{m_2} \times \cdots \times Z_{m_u}$, where t and $m_1, m_2, \ldots, m_u$ satisfy (a) and (b) (i.e. $t \geq 0$, $m_j \geq 2$ for all j and $m_{i+1} \mid m_i$ for $1 \leq i \leq u-1$), then $t = r$, $u = s$ and $m_i = n_i$ for all i.

Proof: We shall derive this theorem in Section 12.1 as a consequence of a more general classification theorem. For finite groups we shall give an alternate proof at the end of Section 6.1.

Definition. The integer r in Theorem 3 is called the *free rank* or *Betti number* of G and the integers $n_1, n_2, \ldots, n_s$ are called the *invariant factors* of G. The description of G in Theorem 3(1) is called the *invariant factor decomposition* of G.

Theorem 3 asserts that the free rank and (ordered) list of invariant factors of an abelian group are uniquely determined, so that two finitely generated abelian groups are isomorphic if and only if they have the same free rank and the same list of invariant factors. Observe that a finitely generated abelian group is a finite group if and only if its free rank is zero.

The order of a finite abelian group is just the product of its invariant factors (by Proposition 1). If G is a finite abelian group with invariant factors $n_1, n_2, \ldots, n_s$, where $n_{i+1} \mid n_i$, $1 \leq i \leq s-1$, then G is said to be of *type* $(n_1, n_2, \ldots, n_s)$.

Theorem 3 gives an effective way of listing *all* finite abelian groups of a given order. Namely, to find (up to isomorphism) all abelian groups of a given order n one must find all finite sequences of integers $n_1, n_2, \ldots, n_s$ such that

(1) $n_j \geq 2$ for all $j \in \{1, 2, \ldots, s\}$,
(2) $n_{i+1} \mid n_i$, $1 \leq i \leq s-1$, and
(3) $n_1 n_2 \cdots n_s = n$.

Theorem 3 states that there is a bijection between the set of such sequences and the set of isomorphism classes of finite abelian groups of order n (where each sequence corresponds to the list of invariant factors of a finite abelian group).

Before illustrating how to find all such sequences for a specific value of n we make some general comments. First note that $n_1 \geq n_2 \geq \cdots \geq n_s$, so n_1 is the largest invariant factor. Also, by property (3) each n_i divides n. If p is any prime divisor of n then by (3) we see that p must divide n_i for some i. Then, by (2), p also divides n_j for all $j \leq i$. It follows that

every prime divisor of n must divide the first invariant factor n_1.

In particular, if n is the product of distinct primes (all to the first power)[1] we see that $n \mid n_1$, hence $n = n_1$. This proves that if n is square free, there is only one possible list of invariant factors for an abelian group of order n (namely, the list $n_1 = n$):

Corollary 4. If n is the product of distinct primes, then up to isomorphism the only abelian group of order n is the cyclic group of order n, Z_n.

The factorization of n into prime powers is the first step in determining all possible lists of invariant factors for abelian groups of order n.

Example

Suppose $n = 180 = 2^2 \cdot 3^2 \cdot 5$. As noted above we must have $2 \cdot 3 \cdot 5 \mid n_1$, so possible values of n_1 are:

$$n_1 = 2^2 \cdot 3^2 \cdot 5, \quad 2^2 \cdot 3 \cdot 5, \quad 2 \cdot 3^2 \cdot 5, \quad \text{or} \quad 2 \cdot 3 \cdot 5.$$

For each of these one must work out all possible n_2's (subject to $n_2 \mid n_1$ and $n_1 n_2 \mid n$). For each resulting pair n_1, n_2 one must work out all possible n_3's etc. until all lists satisfying (1) to (3) are obtained.

[1] Such integers are called *square free* since they are not divisible by any square > 1.

For instance, if $n_1 = 2 \cdot 3^2 \cdot 5$, the only number n_2 dividing n_1 with $n_1 n_2$ dividing n is $n_2 = 2$. In this case $n_1 n_2 = n$, so this list is complete: $2 \cdot 3^2 \cdot 5,\ 2$. The abelian group corresponding to this list is: $Z_{90} \times Z_2$.

If $n_1 = 2 \cdot 3 \cdot 5$, the only candidates for n_2 are $n_2 = 2, 3$ or 6. If $n_2 = 2$ or 3, then since $n_3 \mid n_2$ we would necessarily have $n_3 = n_2$ (and there must be a third term in the list by property (3)). This leads to a contradiction because $n_1 n_2 n_3$ would be divisible by 2^3 or 3^3 respectively, but n is not divisible by either of these numbers. Thus the only list of invariant factors whose first term is $2 \cdot 3 \cdot 5$ is $2 \cdot 3 \cdot 5,\ 2 \cdot 3$. The corresponding abelian group is $Z_{30} \times Z_6$.

Similarly, all permissible lists of invariant factors and the corresponding abelian groups of order 180 are easily seen to be the following:

Invariant Factors	Abelian Groups
$2^2 \cdot 3^2 \cdot 5$	Z_{180}
$2 \cdot 3^2 \cdot 5, 2$	$Z_{90} \times Z_2$
$2^2 \cdot 3 \cdot 5, 3$	$Z_{60} \times Z_3$
$2 \cdot 3 \cdot 5, 2 \cdot 3$	$Z_{30} \times Z_6$

The process we carried out above was somewhat *ad hoc*, however it indicates that the determination of lists of invariant factors of all abelian groups of a given order n relies strongly on the factorization of n. The following theorem (which we shall see is equivalent to the Fundamental Theorem in the case of finite abelian groups) gives a more systematic and computationally much faster way of determining all finite abelian groups of a given order. More specifically, if the factorization of n is

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k},$$

it shows that all permissible lists of invariant factors for abelian groups of order n may be determined by finding permissible lists for groups of order $p_i^{\alpha_i}$ for each i. For a prime power, p^α, we shall see that the problem of determining all permissible lists is equivalent to the determination of all partitions of α (and does not depend on p).

Theorem 5. Let G be an abelian group of order $n > 1$ and let the unique factorization of n into distinct prime powers be

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}.$$

Then

(1) $G \cong A_1 \times A_2 \times \cdots \times A_k$, where $|A_i| = p_i^{\alpha_i}$

(2) for each $A \in \{A_1, A_2, \ldots, A_k\}$ with $|A| = p^\alpha$,

$$A \cong Z_{p^{\beta_1}} \times Z_{p^{\beta_2}} \times \cdots \times Z_{p^{\beta_t}}$$

with $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_t \geq 1$ and $\beta_1 + \beta_2 + \cdots + \beta_t = \alpha$ (where t and $\beta_1, \ldots, \beta_t$ depend on i)

(3) the decomposition in (1) and (2) is unique, i.e. if $G \cong B_1 \times B_2 \times \cdots \times B_m$, with $|B_i| = p_i^{\alpha_i}$ for all i, then $B_i \cong A_i$ and B_i and A_i have the same invariant factors.

Definition. The integers p^{β_j} described in the preceding theorem are called the *elementary divisors* of G. The description of G in Theorem 5(1) and 5(2) is called the *elementary divisor decomposition* of G.

The subgroups A_i described in part (1) of the theorem are the Sylow p_i-subgroups of G. Thus (1) says that G is isomorphic to the direct product of its Sylow subgroups (note that they are normal – since G is abelian – hence unique). Part (1) is often referred to as *The Primary Decomposition Theorem* for finite abelian groups.[2] As with Theorem 3, we shall prove this theorem later.

Note that for p a prime, $p^\beta \mid p^\gamma$ if and only if $\beta \leq \gamma$. Furthermore, $p^{\beta_1} \cdots p^{\beta_t} = p^\alpha$ if and only if $\beta_1 + \cdots + \beta_t = \alpha$. Thus the decomposition of A appearing in part (2) of Theorem 5 is the invariant factor decomposition of A with the "divisibility" conditions on the integers p^{β_j} translated into "additive" conditions on their exponents. The *elementary divisors* of G are now seen to be the *invariant factors of the Sylow p-subgroups* as p runs over all prime divisors of G.

By Theorem 5, in order to find all abelian groups of order $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ one must find for each i, $1 \leq i \leq k$, all possible lists of invariant factors for groups of order $p_i^{\alpha_i}$. The set of elementary divisors of each abelian group is then obtained by taking one set of invariant factors from each of the k lists. The abelian groups are the direct products of the cyclic groups whose orders are the elementary divisors (and distinct lists of elementary divisors give nonisomorphic groups). The advantage of this process over the one described following Theorem 2 is that it is easier to systematize how to obtain all possible lists of invariant factors, $p^{\beta_1}, p^{\beta_2}, \ldots, p^{\beta_t}$, for a group of prime power order p^β. Conditions (1) to (3) for invariant factors described earlier then become

(1) $\beta_j \geq 1$ for all $j \in \{1, 2, \ldots, t\}$,
(2) $\beta_i \geq \beta_{i+1}$ for all i, and
(3) $\beta_1 + \beta_2 + \cdots + \beta_t = \beta$.

Hence, each list of invariant factors in this case is simply a *partition* of β (ordered in descending order). In particular, the number of nonisomorphic abelian groups of order p^β (= the number of distinct lists) equals the number of partitions of β. This number is independent of the prime p. For example the number of abelian groups of order p^5 is obtained from the list of partitions of 5:

Partition of 5	Abelian Groups
5	Z_{p^5}
4, 1	$Z_{p^4} \times Z_p$
3, 2	$Z_{p^3} \times Z_{p^2}$
3, 1, 1	$Z_{p^3} \times Z_p \times Z_p$
2, 2, 1	$Z_{p^2} \times Z_{p^2} \times Z_p$
2, 1, 1, 1	$Z_{p^2} \times Z_p \times Z_p \times Z_p$
1, 1, 1, 1, 1	$Z_p \times Z_p \times Z_p \times Z_p \times Z_p$

[2] Recall that for abelian groups the Sylow p-subgroups are sometimes called the p-primary components.

Thus there are precisely 7 nonisomorphic groups of order p^5, the first in the list being the cyclic group, Z_{p^5}, and the last in the list being the elementary abelian group, E_{p^5}.

If $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ and q_i is the number of partitions of α_i, we see that the number of (distinct, nonisomorphic) abelian groups of order n equals $q_1 q_2 \cdots q_k$.

Example

If $n = 1800 = 2^3 3^2 5^2$ we list the abelian groups of this order as follows:

Order p^β	Partitions of β	Abelian Groups
2^3	3; 2, 1; 1, 1, 1	$Z_8, Z_4 \times Z_2, Z_2 \times Z_2 \times Z_2$
3^2	2; 1, 1	$Z_9, Z_3 \times Z_3$
5^2	2; 1, 1	$Z_{25}, Z_5 \times Z_5$

We obtain the abelian groups of order 1800 by taking one abelian group from each of the three lists (right hand column above) and taking their direct product. Doing this in all possible ways gives all isomorphism types:

$Z_8 \times Z_9 \times Z_{25}$
$Z_8 \times Z_9 \times Z_5 \times Z_5$
$Z_8 \times Z_3 \times Z_3 \times Z_{25}$
$Z_8 \times Z_3 \times Z_3 \times Z_5 \times Z_5$
$Z_4 \times Z_2 \times Z_9 \times Z_{25}$
$Z_4 \times Z_2 \times Z_9 \times Z_5 \times Z_5$
$Z_4 \times Z_2 \times Z_3 \times Z_3 \times Z_{25}$
$Z_4 \times Z_2 \times Z_3 \times Z_3 \times Z_5 \times Z_5$
$Z_2 \times Z_2 \times Z_2 \times Z_9 \times Z_{25}$
$Z_2 \times Z_2 \times Z_2 \times Z_9 \times Z_5 \times Z_5$
$Z_2 \times Z_2 \times Z_2 \times Z_3 \times Z_3 \times Z_{25}$
$Z_2 \times Z_2 \times Z_2 \times Z_3 \times Z_3 \times Z_5 \times Z_5$.

By the Fundamental Theorems above, this is a *complete list* of all abelian groups of order 1800 — every abelian group of this order is isomorphic to precisely one of the groups above and no two of the groups in this list are isomorphic.

We emphasize that the elementary divisors of G are not invariant factors of G (but invariant factors of *subgroups* of G). For instance, in case 1 above the elementary divisors: 8, 9, 25 do not satisfy the divisibility criterion of a list of invariant factors.

Our next aim is to illustrate how to pass from a list of invariant factors of a finite abelian group to its list of elementary divisors and vice versa. We show how to determine these invariants of the group no matter how it is given as a direct product of cyclic groups. We need the following proposition:

Proposition 6. Let $m, n \in \mathbb{Z}^+$.

(1) $Z_m \times Z_n \cong Z_{mn}$ if and only if $(m, n) = 1$.

(2) If $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ then $Z_n \cong Z_{p_1^{\alpha_1}} \times Z_{p_2^{\alpha_2}} \times \cdots \times Z_{p_k^{\alpha_k}}$.

Proof: Since (2) is an easy exercise using (1) and induction on k, we concentrate on proving (1). Let $Z_m = \langle x \rangle$, $Z_n = \langle y \rangle$ and let $l = \text{l.c.m.}(m, n)$. Note that $l = mn$ if and only if $(m, n) = 1$. Let $x^a y^b$ be a typical element of $Z_m \times Z_n$. Then (as noted in Example 1, Section 1)

$$\begin{aligned}(x^a y^b)^l &= x^{la} y^{lb} \\ &= 1^a 1^b = 1 \qquad (\text{because } m \mid l \text{ and } n \mid l).\end{aligned}$$

If $(m, n) \neq 1$, every element of $Z_m \times Z_n$ has order at most l, hence has order strictly less than mn, so $Z_m \times Z_n$ cannot be isomorphic to Z_{mn}.

Conversely, if $(m, n) = 1$, then $|xy| = \text{l.c.m.}(|x|, |y|) = mn$. Thus, by order considerations, $Z_m \times Z_n = \langle xy \rangle$ is cyclic, completing the proof.

Obtaining Elementary Divisors from Invariant Factors:

Suppose G is given as an abelian group of type $(n_1, n_2, \ldots, n_s)$, that is

$$G \cong Z_{n_1} \times Z_{n_2} \times \cdots \times Z_{n_s}.$$

Let $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k} = n_1 n_2 \cdots n_s$. Factor each n_i as

$$n_i = p_1^{\beta_{i1}} p_2^{\beta_{i2}} \cdots p_k^{\beta_{ik}}, \quad \text{where } \beta_{ij} \geq 0.$$

By the proposition above,

$$Z_{n_i} \cong Z_{p_1^{\beta_{i1}}} \times \cdots \times Z_{p_k^{\beta_{ik}}},$$

for each i. If $\beta_{ij} = 0$, $Z_{p_j^{\beta_{ij}}} = 1$ and this factor may be deleted from the direct product without changing the isomorphism type. Then the elementary divisors of G are precisely the integers

$$p_j^{\beta_{ij}}, \qquad 1 \leq j \leq k, \quad 1 \leq i \leq s \text{ such that } \beta_{ij} \neq 0.$$

For example, if $|G| = 2^3 \cdot 3^2 \cdot 5^2$ and G is of type $(30, 30, 2)$, then

$$G \cong Z_{30} \times Z_{30} \times Z_2.$$

Since $Z_{30} \cong Z_2 \times Z_3 \times Z_5$, $G \cong Z_2 \times Z_3 \times Z_5 \times Z_2 \times Z_3 \times Z_5 \times Z_2$. The elementary divisors of G are therefore 2, 3, 5, 2, 3, 5, 2, or, grouping like primes together (note that rearranging the order of the factors in a direct product does not affect the isomorphism type (Exercise 7 of Section 1)), 2, 2, 2, 3, 3, 5, 5. In particular, G is isomorphic to the last group in the list in the example above.

If for each j one collects all the factors $Z_{p_j^{\beta_{ij}}}$ together, the resulting direct product forms the Sylow p_j-subgroup, A_j, of G. Thus the Sylow 2-subgroup of the group in the preceding paragraph is isomorphic to $Z_2 \times Z_2 \times Z_2$ (i.e. the elementary abelian group of order 8).

Obtaining Elementary Divisors from any Cyclic Decomposition:

The same process described above will give the elementary divisors of a finite abelian group G whenever G is given as a direct product of cyclic groups (not just when the orders of the cyclic components are the invariant factors). For example, if $G = Z_6 \times Z_{15}$, the list 6, 15 is neither that of the invariant factors (the divisibility condition fails) nor that of elementary divisors (they are not prime powers). To find the elementary divisors, factor $6 = 2 \cdot 3$ and $15 = 3 \cdot 5$. Then the prime powers 2, 3, 3, 5 are the elementary divisors and

$$G \cong Z_2 \times Z_3 \times Z_3 \times Z_5.$$

Obtaining Invariant Factors from Elementary Divisors:

Suppose G is an abelian group of order n, where $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ and we are given the elementary divisors of G. The invariant factors of G are obtained by following these steps:

(1) First group all elementary divisors which are powers of the same prime together. In this way we obtain k lists of integers (one for each p_j).

(2) In each of these k lists arrange the integers in nonincreasing order.

(3) Among these k lists suppose that the longest (i.e. the one with the most terms) consists of t integers. Make each of the k lists of length t by appending an appropriate number of 1's at the end of each list.

(4) For each $i \in \{1, 2, \ldots, t\}$ the ith invariant factor, n_i, is obtained by taking the product of the ith integer in each of the t (ordered) lists.

The point of ordering the lists in this way is to ensure that we have the divisibility condition $n_{i+1} \mid n_i$.

Suppose, for example, that the elementary divisors of G are given as 2, 3, 2, 25, 3, 2 (so $|G| = 2^3 \cdot 3^2 \cdot 5^2$). Regrouping and increasing each list to have 3 ($= t$) members gives:

$p = 2$	$p = 3$	$p = 5$
2	3	25
2	3	1
2	1	1

so the invariant factors of G are: $2 \cdot 3 \cdot 25$, $\quad 2 \cdot 3 \cdot 1$, $\quad 2 \cdot 1 \cdot 1$ and

$$G \cong Z_{150} \times Z_6 \times Z_2.$$

Note that this is the penultimate group in the list classifying abelian groups of order 1800 computed above.

The invariant factor decompositions of the abelian groups of order 1800 are as follows, where the ith group in this list is isomorphic to the ith group computed in the previous list:

$$
\begin{aligned}
&Z_{1800}\\
&Z_{360} \times Z_5\\
&Z_{600} \times Z_3\\
&Z_{120} \times Z_{15}\\
&Z_{900} \times Z_2\\
&Z_{180} \times Z_{10}\\
&Z_{300} \times Z_6\\
&Z_{60} \times Z_{30}\\
&Z_{450} \times Z_2 \times Z_2\\
&Z_{90} \times Z_{10} \times Z_2\\
&Z_{150} \times Z_6 \times Z_2\\
&Z_{30} \times Z_{30} \times Z_2.
\end{aligned}
$$

Using the uniqueness statements of the Fundamental Theorems 3 and 5, we can use these processes to determine whether any two direct products of finite cyclic groups are isomorphic. For instance, if one wanted to know whether $Z_6 \times Z_{15} \cong Z_{10} \times Z_9$, first determine whether they have the same order (both are of order 90) and then (the easiest way in general) determine whether they have the same elementary divisors:

$Z_6 \times Z_{15}$ has elementary divisors 2, 3, 3, 5 and is isomorphic to $Z_2 \times Z_3 \times Z_3 \times Z_5$

$Z_{10} \times Z_9$ has elementary divisors 2, 5, 9 and is isomorphic to $Z_2 \times Z_5 \times Z_9$.

The lists of elementary divisors are different so (by Theorem 5) they are not isomorphic. Note that $Z_6 \times Z_{15}$ has no element of order 9 whereas $Z_{10} \times Z_9$ does (cf. Exercise 5).

The processes we described above (with some elaboration) form a proof (via Proposition 6) that for finite abelian groups Theorems 3 and 5 are equivalent (i.e. one implies the other). We leave the details to the reader.

One can now better understand some of the power and some of the limitations of classification theorems. On one hand, given any positive integer n one can explicitly describe all abelian groups of order n, a significant achievement. On the other hand, the amount of information necessary to determine which of the isomorphism types of groups of order n a particular group belongs to may be considerable (and is large if n is divisible by large powers of primes).

We close this section with some terminology which will be useful in later sections.

Definition

(1) If G is a finite abelian group of type $(n_1, n_2, \ldots, n_t)$, the integer t is called the *rank* of G (the free rank of G is 0 so there will be no confusion).

(2) If G is any group, the *exponent* of G is the smallest positive integer n such that $x^n = 1$ for all $x \in G$ (if no such integer exists the exponent of G is ∞).

EXERCISES

1. In each of parts (a) to (e) give the number of nonisomorphic abelian groups of the specified order — do not list the groups: (a) order 100, (b) order 576, (c) order 1155, (d) order 42875, (e) order 2704.
2. In each of parts (a) to (e) give the lists of invariant factors for all abelian groups of the specified order:
(a) order 270, (b) order 9801, (c) order 320, (d) order 105, (e) order 44100.
3. In each of parts (a) to (e) give the lists of elementary divisors for all abelian groups of the specified order and then match each list with the corresponding list of invariant factors found in the preceding exercise:
(a) order 270, (b) order 9801, (c) order 320, (d) order 105, (e) order 44100.
4. In each of parts (a) to (d) determine which pairs of abelian groups listed are isomorphic (here the expression $\{a_1, a_2, \ldots, a_k\}$ denotes the abelian group $Z_{a_1} \times Z_{a_2} \times \cdots \times Z_{a_k}$).
(a) $\{4, 9\}$, $\{6, 6\}$, $\{8, 3\}$, $\{9, 4\}$, $\{6, 4\}$, $\{64\}$.
(b) $\{2^2, 2 \cdot 3^2\}$, $\{2^2 \cdot 3, 2 \cdot 3\}$, $\{2^3 \cdot 3^2\}$, $\{2^2 \cdot 3^2, 2\}$.
(c) $\{5^2 \cdot 7^2, 3^2 \cdot 5 \cdot 7\}$, $\{3^2 \cdot 5^2 \cdot 7, 5 \cdot 7^2\}$, $\{3 \cdot 5^2, 7^2, 3 \cdot 5 \cdot 7\}$, $\{5^2 \cdot 7, 3^2 \cdot 5, 7^2\}$.
(d) $\{2^2 \cdot 5 \cdot 7, 2^3 \cdot 5^3, 2 \cdot 5^2\}$, $\{2^3 \cdot 5^3 \cdot 7, 2^3 \cdot 5^3\}$, $\{2^2, 2 \cdot 7, 2^3, 5^3, 5^3\}$, $\{2 \cdot 5^3, 2^2 \cdot 5^3, 2^3, 7\}$.
5. Let G be a finite abelian group of type $(n_1, n_2, \ldots, n_t)$. Prove that G contains an element of order m if and only if $m \mid n_1$. Deduce that G is of exponent n_1.
6. Prove that any finite group has a finite exponent. Give an example of an infinite group with finite exponent. Does a finite group of exponent m always contain an element of order m?
7. Let p be a prime and let $A = \langle x_1 \rangle \times \langle x_2 \rangle \times \cdots \times \langle x_n \rangle$ be an abelian p-group, where $|x_i| = p^{\alpha_i} > 1$ for all i. Define the pth*-power map*

$$\varphi : A \to A \qquad \text{by} \qquad \varphi : x \mapsto x^p.$$

(a) Prove that φ is a homomorphism.
(b) Describe the image and kernel of φ in terms of the given generators.
(c) Prove that both $\ker \varphi$ and $A/\operatorname{im} \varphi$ have rank n (i.e. have the same rank as A) and prove that these groups are both isomorphic to the elementary abelian group of order p^n: E_{p^n}.
8. Let A be a finite abelian group (written multiplicatively) and let p be a prime. Let

$$A^p = \{a^p \mid a \in A\} \qquad \text{and} \qquad A_p = \{x \mid x^p = 1\}$$

(so A^p and A_p are the image and kernel of the pth-power map, respectively).
(a) Prove that $A/A^p \cong A_p$. [Show that they are both elementary abelian and they have the same order.]
(b) Prove that the number of subgroups of A of order p equals the number of subgroups of A of index p. [Reduce to the case where A is an elementary abelian p-group.]
9. Let $A = Z_{60} \times Z_{45} \times Z_{12} \times Z_{36}$. Find the number of elements of order 2 and the number of subgroups of index 2 in A.

10. Let n and k be positive integers and let A be the free abelian group of rank n (written additively). Prove that A/kA is isomorphic to the direct product of n copies of $\mathbb{Z}/k\mathbb{Z}$ (here $kA = \{ka \mid a \in A\}$). [See Exercise 14, Section 1.]

11. Let G be a nontrivial finite abelian group of rank t.
    (a) Prove that the rank of G equals the maximum of the ranks of its Sylow subgroups.
    (b) Prove that G can be generated by t elements but no subset with fewer than t elements generates G. [One way of doing this is by using part (a) together with Exercise 7.]

12. Let n and m be positive integers with $d = (n, m)$. Let $Z_n = \langle x \rangle$ and $Z_m = \langle y \rangle$. Let A be the central product of $\langle x \rangle$ and $\langle y \rangle$ with an element of order d identified: $\langle x, y \mid x^n = y^m = 1,\ xy = yx,\ x^{\frac{n}{d}} = y^{\frac{m}{d}} \rangle$. Describe A as a direct product of two cyclic groups.

13. Let $A = \langle x_1 \rangle \times \cdots \times \langle x_r \rangle$ be a finite abelian group with $|x_i| = n_i$ for $1 \le i \le r$. Find a presentation for A. Prove that if G is any group containing commuting elements $g_1, \ldots, g_r$ such that $g_i^{n_i} = 1$ for $1 \le i \le r$, then there is a unique homomorphism from A to G which sends x_i to g_i for all i.

14. For any group G define the *dual group* of G (denoted $\widehat{G}$) to be the set of all homomorphisms from G into the multiplicative group of roots of unity in $\mathbb{C}$. Define a group operation in $\widehat{G}$ by pointwise multiplication of functions: if χ, ψ are homomorphisms from G into the group of roots of unity then $\chi\psi$ is the homomorphism given by $(\chi\psi)(g) = \chi(g)\psi(g)$ for all $g \in G$, where the latter multiplication takes place in $\mathbb{C}$.
    (a) Show that this operation on $\widehat{G}$ makes $\widehat{G}$ into an abelian group. [Show that the identity is the map $g \mapsto 1$ for all $g \in G$ and the inverse of $\chi \in \widehat{G}$ is the map $g \mapsto \chi(g)^{-1}$.]
    (b) If G is a finite abelian group, prove that $\widehat{G} \cong G$. [Write G as $\langle x_1 \rangle \times \cdots \times \langle x_r \rangle$ and if $n_i = |x_i|$ define χ_i to be the homomorphism which sends x_i to $e^{2\pi i/n_i}$ and sends x_j to 1, for all $j \ne i$. Prove χ_i has order n_i in $\widehat{G}$ and $\widehat{G} = \langle \chi_1 \rangle \times \cdots \times \langle \chi_r \rangle$.]

    (This result is often phrased: a finite abelian group is self-dual. It implies that the lattice diagram of a finite abelian group is the same when it is turned upside down. Note however that there is no *natural* isomorphism between G and its dual (the isomorphism depends on a choice of a set of generators for G). This is frequently stated in the form: a finite abelian group is *noncanonically* isomorphic to its dual.)

15. Let $G = \langle x \rangle \times \langle y \rangle$ where $|x| = 8$ and $|y| = 4$.
    (a) Find all pairs a, b in G such that $G = \langle a \rangle \times \langle b \rangle$ (where a and b are expressed in terms of x and y).
    (b) Let $H = \langle x^2y,\ y^2 \rangle \cong Z_4 \times Z_2$. Prove that there are no elements a, b of G such that $G = \langle a \rangle \times \langle b \rangle$ and $H = \langle a^2 \rangle \times \langle b^2 \rangle$ (i.e. one cannot pick direct product generators for G in such a way that some powers of these are direct product generators for H).

16. Prove that no finitely generated abelian group is divisible (cf. Exercise 19, Section 2.4).

5.3 TABLE OF GROUPS OF SMALL ORDER

At this point we can give in the following table the isomorphism types for most of the groups of small order.

Each of the unfamiliar non-abelian groups in the table will be constructed in Section 5 below on semidirect products (which will also explain the notation used for them). For the present we give generators and relations for each of them (i.e. presentations of them).

Order	No. of Isomorphism Types	Abelian Groups	Non-abelian Groups
1	1	Z_1	none
2	1	Z_2	none
3	1	Z_3	none
4	2	$Z_4, Z_2 \times Z_2$	none
5	1	Z_5	none
6	2	Z_6	S_3
7	1	Z_7	none
8	5	$Z_8, Z_4 \times Z_2,$ $Z_2 \times Z_2 \times Z_2$	D_8, Q_8
9	2	$Z_9, Z_3 \times Z_3$	none
10	2	Z_{10}	D_{10}
11	1	Z_{11}	none
12	5	$Z_{12}, Z_6 \times Z_2$	$A_4, D_{12}, Z_3 \rtimes Z_4$
13	1	Z_{13}	none
14	2	Z_{14}	D_{14}
15	1	Z_{15}	none
16	14	$Z_{16}, Z_8 \times Z_2,$ $Z_4 \times Z_4, Z_4 \times Z_2 \times Z_2,$ $Z_2 \times Z_2 \times Z_2 \times Z_2$	not listed
17	1	Z_{17}	none
18	5	Z_{18}, Z_6, Z_3	$D_{18}, S_3 \times Z_3,$ $(Z_3 \times Z_3) \rtimes Z_2$
19	1	Z_{19}	none
20	5	$Z_{20}, Z_{10} \times Z_2$	D_{20} $Z_5 \rtimes Z_4, F_{20}$

The group $Z_3 \rtimes Z_4$ of order 12 can be described by the generators and relations:

$$\langle x, y \mid x^4 = y^3 = 1,\ x^{-1}yx = y^{-1} \rangle,$$

namely, it has a normal Sylow 3-subgroup ($\langle y \rangle$) which is inverted by an element of order 4 (x) acting by conjugation (x^2 centralizes y).

The group $(Z_3 \times Z_3) \rtimes Z_2$ has generators and relations:

$$\langle x, y, z \mid x^2 = y^3 = z^3 = 1,\ yz = zy,\ x^{-1}yx = y^{-1},\ x^{-1}zx = z^{-1} \rangle,$$

namely, it has a normal Sylow 3-subgroup isomorphic to $Z_3 \times Z_3$ ($\langle y, z \rangle$) inverted by an element of order 2 (x) acting by conjugation.

The group $Z_5 \rtimes Z_4$ of order 20 has generators and relations:

$$\langle x, y \mid x^4 = y^5 = 1,\ x^{-1}yx = y^{-1} \rangle,$$

namely, it has a normal Sylow 5-subgroup ($\langle y \rangle$) which is inverted by an element of order 4 (x) acting by conjugation (x^2 centralizes y).

The group F_{20} of order 20 has generators and relations:

$$\langle x, y \mid x^4 = y^5 = 1,\ x^{-1}yx = y^2 \rangle,$$

namely, it has a normal Sylow 5-subgroup ($\langle y \rangle$) which is squared by an element of order 4 (x) acting by conjugation. One can check that this group occurs as the normalizer of a Sylow 5-subgroup in S_5, e.g.

$$F_{20} = \langle (2\,3\,5\,4),\ (1\,2\,3\,4\,5) \rangle.$$

This group is called the *Frobenius group* of order 20.

EXERCISE

1. Prove that D_{16}, $Z_2 \times D_8$, $Z_2 \times Q_8$, $Z_4 * D_8$, QD_{16} and M are nonisomorphic non-abelian groups of order 16 (where $Z_4 * D_8$ is described in Exercise 12, Section 1 and QD_{16} and M are described in the exercises in Section 2.5).

5.4 RECOGNIZING DIRECT PRODUCTS

So far we have seen that direct products may be used to both construct "larger" groups from "smaller" ones and to decompose finitely generated abelian groups into cyclic factors. Even certain non-abelian groups, which may be given in some other form, may be decomposed as direct products of smaller groups. The purpose of this section is to indicate a criterion to recognize when a group is the direct product of some of its subgroups and to illustrate the criterion with some examples.

Before doing so we introduce some standard notation and elementary results on commutators which will streamline the presentation and which will be used again in Chapter 6 when we consider nilpotent groups.

Definition. Let G be a group, let $x, y \in G$ and let A, B be nonempty subsets of G.

(1) Define $[x, y] = x^{-1}y^{-1}xy$, called the *commutator* of x and y.

(2) Define $[A, B] = \langle [a, b] \mid a \in A,\ b \in B \rangle$, the group generated by commutators of elements from A and from B.

(3) Define $G' = \langle [x, y] \mid x, y \in G \rangle$, the subgroup of G generated by commutators of elements from G, called the *commutator subgroup* of G.

The commutator of x and y is 1 if and only if x and y commute, which explains the terminology. The following proposition shows how commutators measure the "difference" in G between xy and yx.

Proposition 7. Let G be a group, let $x, y \in G$ and let $H \leq G$. Then

(1) $xy = yx[x, y]$ (in particular, $xy = yx$ if and only if $[x, y] = 1$).

(2) $H \trianglelefteq G$ if and only if $[H, G] \leq H$.

(3) $\sigma[x, y] = [\sigma(x), \sigma(y)]$ for any automorphism σ of G, G' char G and G/G' is abelian.

(4) G/G' is the largest abelian quotient of G in the sense that if $H \trianglelefteq G$ and G/H is abelian, then $G' \leq H$. Conversely, if $G' \leq H$, then $H \trianglelefteq G$ and G/H is abelian.

(5) If $\varphi : G \to A$ is any homomorphism of G into an abelian group A, then φ factors through G' i.e. $G' \leq \ker \varphi$ and the following diagram commutes:

$$\begin{tikzcd} G \arrow[r] \arrow[dr, "\varphi"'] & G/G' \arrow[d] \\ & A \end{tikzcd}$$

Proof: (1) This is immediate from the definition of $[x, y]$.

(2) By definition, $H \trianglelefteq G$ if and only if $g^{-1}hg \in H$ for all $g \in G$ and all $h \in H$. For $h \in H$, $g^{-1}hg \in H$ if and only if $h^{-1}g^{-1}hg \in H$, so that $H \trianglelefteq G$ if and only if $[h, g] \in H$ for all $h \in H$ and all $g \in G$. Thus $H \trianglelefteq G$ if and only if $[H, G] \leq H$, which is (2).

(3) Let $\sigma \in \mathrm{Aut}(G)$ be an automorphism of G and let $x, y \in G$. Then

$$\begin{aligned} \sigma([x, y]) &= \sigma(x^{-1}y^{-1}xy) \\ &= \sigma(x)^{-1}\sigma(y)^{-1}\sigma(x)\sigma(y) \\ &= [\sigma(x), \sigma(y)]. \end{aligned}$$

Thus for every commutator $[x, y]$ of G', $\sigma([x, y])$ is again a commutator. Since σ has a 2-sided inverse, it follows that it maps the set of commutators bijectively onto itself. Since the commutators are a generating set for G', $\sigma(G') = G'$, that is, G' char G.

To see that G/G' is abelian, let xG' and yG' be arbitrary elements of G/G'. By definition of the group operation in G/G' and since $[x, y] \in G'$ we have

$$\begin{aligned} (xG')(yG') &= (xy)G' \\ &= (yx[x, y])G' \\ &= (yx)G' = (yG')(xG'), \end{aligned}$$

which completes the proof of (3).

(4) Suppose $H \trianglelefteq G$ and G/H is abelian. Then for all $x, y \in G$, $(xH)(yH) = (yH)(xH)$ so

$$\begin{aligned} 1H &= (xH)^{-1}(yH)^{-1}(xH)(yH) \\ &= x^{-1}y^{-1}xyH \\ &= [x, y]H. \end{aligned}$$

Thus $[x, y] \in H$ for all $x, y \in G$, so that $G' \leq H$.

Conversely, if $G' \leq H$, then since G/G' is abelian by (3), every subgroup of G/G' is normal. In particular, $H/G' \trianglelefteq G/G'$. By the Lattice Isomorphism Theorem $H \trianglelefteq G$. By the Third Isomorphism Theorem

$$G/H \cong (G/G')/(H/G')$$

hence G/H is abelian (being isomorphic to a quotient of the abelian group G/G'). This proves (4).

(5) This is (4) phrased in terms of homomorphisms.

Passing to the quotient by the commutator subgroup of G collapses all commutators to the identity so that all elements in the quotient group commute. As (4) indicates, a strong converse to this also holds: a quotient of G by H is abelian if and only if the commutator subgroup is contained in H (i.e. if and only if G' is mapped to the identity in the quotient G/H).

We shall exhibit a group (of order 96) in the next section with the property that one of the elements of its commutator subgroup *cannot* be written as a single commutator $[x, y]$ for any x and y. Thus G' does not necessarily consist only of the set of (single) commutators (but is the group *generated* by these elements).

Examples

(1) A group G is abelian if and only if $G' = 1$.

(2) Sometimes it is possible to compute the commutator subgroup of a group without actually calculating commutators explicitly. For instance, if $G = D_8$, then since $Z(D_8) = \langle r^2 \rangle \trianglelefteq D_8$ and $D_8/Z(D_8)$ is abelian (the Klein 4-group), the commutator subgroup D_8' is a subgroup of $Z(D_8)$. Since D_8 is not itself abelian its commutator subgroup is nontrivial. The only possibility is that $D_8' = Z(D_8)$. By a similar argument, $Q_8' = Z(Q_8) = \langle -1 \rangle$. More generally, if G is any non-abelian group of order p^3, where p is a prime, $G' = Z(G)$ and $|G'| = p$ (Exercise 7).

(3) Let $D_{2n} = \langle r, s \mid r^n = s^2 = 1, s^{-1}rs = r^{-1} \rangle$. Then $[r, s] = r^{-2}$, so $\langle r^{-2} \rangle = \langle r^2 \rangle \leq D_{2n}'$. Furthermore, $\langle r^2 \rangle \trianglelefteq D_{2n}$ and the images of r and s in $D_{2n}/\langle r^2 \rangle$ generate this quotient. They are commuting elements of order ≤ 2, so the quotient is abelian and $D_{2n}' \leq \langle r^2 \rangle$. Thus $D_{2n}' = \langle r^2 \rangle$. Finally, note that if n ($= |r|$) is odd, $\langle r^2 \rangle = \langle r \rangle$ whereas if n is even, $\langle r^2 \rangle$ is of index 2 in $\langle r \rangle$. Hence D_{2n}' is of index 2 or 4 in D_{2n} according to whether n is odd or even, respectively.

(4) Since conjugation by $g \in G$ is an automorphism of G, $[a^g, b^g] = [a, b]^g$ for all $a, b \in G$ by (3) of the proposition, i.e. conjugates of commutators are also commutators. For example, once we exhibit an element of one cycle type in S_n as a commutator, every element of the same cycle type is also a commutator (cf. Section 4.3). For example, every 5-cycle is a commutator in S_5 as follows: labelling the vertices of a pentagon as $1, \ldots, 5$ we see that $D_{10} \leq S_5$ (a subgroup of A_5 in fact). By the preceding example an element of order 5 is a commutator in D_{10}, hence also in S_5. Explicitly, $(1\,4\,2\,5\,3) = [(1\,2\,3\,4\,5), (2\,5)(4\,3)]$.

The next result actually follows from the proof of Proposition 3.13 but we isolate it explicitly for reference:

Proposition 8. Let H and K be subgroups of the group G. The number of distinct ways of writing each element of the set HK in the form hk, for some $h \in H$

and $k \in K$ is $|H \cap K|$. In particular, if $H \cap K = 1$, then each element of HK can be written uniquely as a product hk, for some $h \in H$ and $k \in K$.

Proof: Exercise.

The main result of this section is the following *recognition theorem*:

Theorem 9. Suppose G is a group with subgroups H and K such that
(1) H and K are normal in G, and
(2) $H \cap K = 1$.
Then $HK \cong H \times K$.

Proof: Observe that by hypothesis (1), HK is a subgroup of G (see Corollary 3.15). Let $h \in H$ and let $k \in K$. Since $H \trianglelefteq G$, $k^{-1}hk \in H$, so that $h^{-1}(k^{-1}hk) \in H$. Similarly, $(h^{-1}k^{-1}h)k \in K$. Since $H \cap K = 1$ it follows that $h^{-1}k^{-1}hk = 1$, i.e. $hk = kh$ so that every element of H commutes with every element of K.

By the preceding proposition each element of HK can be written uniquely as a product hk, with $h \in H$ and $k \in K$. Thus the map

$$\begin{aligned} \varphi : HK &\to H \times K \\ hk &\mapsto (h, k) \end{aligned}$$

is well defined. To see that φ is a homomorphism note that if $h_1, h_2 \in H$ and $k_1, k_2 \in K$, then we have seen that h_2 and k_1 commute. Thus

$$(h_1k_1)(h_2k_2) = (h_1h_2)(k_1k_2)$$

and the latter product is the unique way of writing $(h_1k_1)(h_2k_2)$ in the form hk with $h \in H$ and $k \in K$. This shows that

$$\begin{aligned} \varphi(h_1k_1h_2k_2) &= \varphi(h_1h_2k_1k_2) \\ &= (h_1h_2, k_1k_2) \\ &= (h_1, k_1)(h_2, k_2) = \varphi(h_1k_1)\varphi(h_2k_2) \end{aligned}$$

so that φ is a homomorphism. The homomorphism φ is a bijection since the representation of each element of HK as a product of the form hk is unique, which proves that φ is an isomorphism.

Definition. If G is a group and H and K are normal subgroups of G with $H \cap K = 1$, we call HK the *internal direct product* of H and K. We shall (when emphasis is called for) call $H \times K$ the *external direct product* of H and K.

The distinction between internal and external direct product is (by Theorem 9) purely notational: the elements of the internal direct product are written in the form hk, whereas those of the external direct product are written as ordered pairs (h, k). We have in previous instances passed between these. For example, when $Z_n = \langle a \rangle$ and $Z_m = \langle b \rangle$ we wrote $x = (a, 1)$ and $y = (1, b)$ so that every element of $Z_n \times Z_m$ was written in the form $x^r y^s$.

Examples

(1) If n is a positive odd integer, we show $D_{4n} \cong D_{2n} \times Z_2$. To see this let

$$D_{4n} = \langle r, s \mid r^{2n} = s^2 = 1,\ srs = r^{-1} \rangle$$

be the usual presentation of D_{4n}. Let $H = \langle s, r^2 \rangle$ and let $K = \langle r^n \rangle$. Geometrically, if D_{4n} is the group of symmetries of a regular $2n$-gon, H is the group of symmetries of the regular n-gon inscribed in the $2n$-gon by joining vertex $2i$ to vertex $2i+2$, for all i mod $2n$ (and if one lets $r_1 = r^2$, H has the usual presentation of the dihedral group of order $2n$ with generators r_1 and s). Note that $H \trianglelefteq D_{4n}$ (it has index 2). Since $|r| = 2n$, $|r^n| = 2$. Since $srs = r^{-1}$, we have $sr^n s = r^{-n} = r^n$, that is, s centralizes r^n. Since clearly r centralizes r^n, $K \leq Z(D_{4n})$. Thus $K \trianglelefteq D_{4n}$. Finally, $K \not\leq H$ since r^2 has odd order (or because r^n sends vertex i into vertex $i+n$, hence does not preserve the set of even vertices of the $2n$-gon). Thus $H \cap K = 1$ by Lagrange. Theorem 9 now completes the proof.

(2) Let I be a subset of $\{1, 2, \ldots, n\}$ and let G be the setwise stabilizer of I in S_n, i.e.

$$G = \{\sigma \in S_n \mid \sigma(i) \in I \text{ for all } i \in I\}.$$

Let $J = \{1, 2, \ldots, n\} - I$ be the complement of I and note that G is also the setwise stabilizer of J. Let H be the *pointwise* stabilizer of I and let K be the *pointwise* stabilizer of $\{1, 2, \ldots, n\} - I$, i.e.

$$H = \{\sigma \in G \mid \sigma(i) = i \text{ for all } i \in I\}$$
$$K = \{\tau \in G \mid \tau(j) = j \text{ for all } j \in J\}.$$

It is easy to see that H and K are normal subgroups of G (in fact they are kernels of the actions of G on I and J, respectively). Since any element of $H \cap K$ fixes all of $\{1, 2, \ldots, n\}$, we have $H \cap K = 1$. Finally, since every element σ of G stabilizes the sets I and J, each cycle in the cycle decomposition of σ involves only elements of I or only elements of J. Thus σ may be written as a product $\sigma_I \sigma_J$, where $\sigma_I \in H$ and $\sigma_J \in K$. This proves $G = HK$. By Theorem 9, $G \cong H \times K$. Now any permutation of J can be extended to a permutation in S_n by letting it act as the identity on I. These are precisely the permutations in H (and similarly the permutations in K are the permutations of I which are the identity on J), so

$$H \cong S_J \quad K \cong S_I \quad \text{and} \quad G \cong S_m \times S_{n-m},$$

where $m = |J|$ (and, by convention, $S_\emptyset = 1$).

(3) Let $\sigma \in S_n$ and let I be the subset of $\{1, 2, \ldots, n\}$ fixed pointwise by σ:

$$I = \{i \in \{1, 2, \ldots, n\} \mid \sigma(i) = i\}.$$

If $C = C_{S_n}(\sigma)$, then by Exercise 18 of Section 4.3, C stabilizes the set I and its complement J. By the preceding example, C is isomorphic to a subgroup of $H \times K$, where H is the subgroup of all permutations in S_n fixing I pointwise and K is the set of all permutations fixing J pointwise. Note that $\sigma \in H$. Thus each element, α, of C can be written (uniquely) as $\alpha = \alpha_I \alpha_J$, for some $\alpha_I \in H$ and $\alpha_J \in K$. Note further that if τ is any permutation of $\{1, 2, \ldots, n\}$ which fixes each $j \in J$ (i.e. any element of K), then σ and τ commute (since they move no common integers). Thus C contains all such τ, i.e. C contains the subgroup K. This proves that the group C consists of all elements $\alpha_I \alpha_J \in H \times K$ such that α_J is arbitrary in K and α_I

commutes with σ in H:

$$\begin{aligned} C_{S_n}(\sigma) &= C_H(\sigma) \times K \\ &\cong C_{S_j}(\sigma) \times S_I. \end{aligned}$$

In particular, if σ is an m-cycle in S_n,

$$C_{S_n}(\sigma) = \langle \sigma \rangle \times S_{n-m}.$$

The latter group has order $m(n-m)!$, as computed in Section 4.3.

EXERCISES

Let G be a group.

1. Prove that if $x, y \in G$ then $[y, x] = [x, y]^{-1}$. Deduce that for any subsets A and B of G, $[A, B] = [B, A]$ (recall that $[A, B]$ is the *subgroup* of G generated by the commutators $[a, b]$).
2. Prove that a subgroup H of G is normal if and only if $[G, H] \leq H$.
3. Let $a, b, c \in G$. Prove that
   (a) $[a, bc] = [a, c](c^{-1}[a, b]c)$
   (b) $[ab, c] = (b^{-1}[a, c]b)[b, c]$.
4. Find the commutator subgroups of S_4 and A_4.
5. Prove that A_n is the commutator subgroup of S_n for all $n \geq 5$.
6. Exhibit a representative of each cycle type of A_5 as a commutator in S_5.
7. Prove that if p is a prime and P is a non-abelian group of order p^3 then $P' = Z(P)$.
8. Assume $x, y \in G$ and both x and y commute with $[x, y]$. Prove that for all $n \in \mathbb{Z}^+$, $(xy)^n = x^n y^n [y, x]^{\frac{n(n-1)}{2}}$.
9. Prove that if p is an odd prime and P is a group of order p^3 then the pth power map $x \mapsto x^p$ is a homomorphism of P into $Z(P)$. If P is not cyclic, show that the kernel of the pth power map has order p^2 or p^3. Is the squaring map a homomorphism in non-abelian groups of order 8? Where is the oddness of p needed in the above proof? [Use Exercise 8.]
10. Prove that a finite abelian group is the direct product of its Sylow subgroups.
11. Prove that if $G = HK$ where H and K are characteristic subgroups of G with $H \cap K = 1$ then $\text{Aut}(G) \cong \text{Aut}(H) \times \text{Aut}(K)$. Deduce that if G is an abelian group of finite order then $\text{Aut}(G)$ is isomorphic to the direct product of the automorphism groups of its Sylow subgroups.
12. Use Theorem 4.17 to describe the automorphism group of a finite cyclic group.
13. Prove that D_{8n} is not isomorphic to $D_{4n} \times Z_2$.
14. Let F be a field and let $G = \{(a_{ij}) \in GL_n(F) \mid a_{ij} = 0 \text{ if } i > j, \text{ and } a_{11} = a_{22} = \cdots = a_{nn}\}$ (i.e. G is the group of upper triangular matrices all of whose diagonal entries are equal). Prove that $G \cong D \times U$, where D is the group of all nonzero multiples of the identity matrix and U is the group of upper triangular matrices with 1's down the diagonal.
15. If A and B are normal subgroups of G such that G/A and G/B are both abelian, prove that $G/(A \cap B)$ is abelian.
16. Prove that if K is a normal subgroup of G then $K' \trianglelefteq G$.
17. If K is a normal subgroup of G and K is cyclic, prove that $G' \leq C_G(K)$. [Recall that the automorphism group of a cyclic group is abelian.]

18. Let $K_1, K_2, \ldots, K_n$ be non-abelian simple groups and let $G = K_1 \times K_2 \times \cdots \times K_n$. Prove that every normal subgroup of G is of the form G_I for some subset I of $\{1, 2, \ldots, n\}$ (where G_I is defined in Exercise 2 of Section 1). [Show that if $N \trianglelefteq G$ and $x = (a_1, \ldots, a_n) \in N$ with some $a_i \neq 1$, then there is some $g_i \in G_i$ not commuting with a_i. Show $[(1, \ldots, g_i, \ldots, 1), x] \in K_i \cap N$ and deduce $K_i \leq N$.]

19. A group H is called *perfect* if $H' = H$ (i.e. H equals its own commutator subgroup).

(a) Prove that every non-abelian simple group is perfect.

(b) Prove that if H and K are perfect subgroups of a group G then $\langle H, K \rangle$ is also perfect. Extend this to show that the subgroup of G generated by any collection of perfect subgroups is perfect.

(c) Prove that any conjugate of a perfect subgroup is perfect.

(d) Prove that any group G has a unique maximal perfect subgroup and that this subgroup is normal.

5.5 SEMIDIRECT PRODUCTS

In this section we study the "semidirect product" of two groups H and K, which is a generalization of the notion of the direct product of H and K obtained by relaxing the requirement that both H and K be normal. This construction will enable us (in certain circumstances) to build a "larger" group from the groups H and K in such a way that G contains subgroups isomorphic to H and K, respectively, as in the case of direct products. In this case the subgroup H will be normal in G but the subgroup K will not necessarily be normal (as it is for direct products). Thus, for instance, we shall be able to construct non-abelian groups even if H and K are abelian. This construction will allow us to enlarge considerably the set of examples of groups at our disposal. As in the preceding section, we shall then prove a recognition theorem that will enable us to decompose some familiar groups into smaller "factors", from which we shall be able to derive some classification theorems.

By way of motivation suppose we already have a group G containing subgroups H and K such that

(a) $H \trianglelefteq G$ (but K is not necessarily normal in G), and

(b) $H \cap K = 1$.

It is still true that HK is a subgroup of G (Corollary 3.15) and, by Proposition 8, every element of HK can be written uniquely as a product hk, for some $h \in H$ and $k \in K$, i.e. there is a bijection between HK and the collection of ordered pairs (h, k), given by $hk \mapsto (h, k)$ (so the group H appears as the set of elements $(h, 1)$ and K appears as the set of elements $(1, k)$). Given two elements h_1k_1 and h_2k_2 of HK, we first see how to write their product (in G) in the same form:

$$\begin{aligned}(h_1k_1)(h_2k_2) &= h_1k_1h_2(k_1^{-1}k_1)k_2 \\ &= h_1(k_1h_2k_1^{-1})k_1k_2 \\ &= h_3k_3,\end{aligned} \tag{5.1}$$

where $h_3 = h_1(k_1h_2k_1^{-1})$ and $k_3 = k_1k_2$. Note that since $H \trianglelefteq G$, $k_1h_2k_1^{-1} \in H$, so $h_3 \in H$ and $k_3 \in K$.

These calculations were predicated on the assumption that there *already existed* a group G containing subgroups H and K with $H \trianglelefteq G$ and $H \cap K = 1$. The basic idea of the semidirect product is to turn this construction around, namely start with two (abstract) groups H and K and try to *define* a group containing (an isomorphic copy of) them in such a way that (a) and (b) above hold. To do this, we write equation (1), which defines the multiplication of elements in our group, in a way that makes sense even if we do not already know there is a group containing H and K as above. The point is that k_3 in equation (1) is obtained only from multiplication in K (namely k_1k_2) and h_3 is obtained from multiplying h_1 and $k_1h_2k_1^{-1}$ in H. If we can understand where the element $k_1h_2k_1^{-1}$ arises (in terms of H and K and without reference to G), then the group HK will have been described entirely in terms of H and K. We can then use this description to *define* the group HK using equation (1) to define the multiplication.

Since H is normal in G, the group K *acts* on H by conjugation:

$$k \bullet h = khk^{-1} \qquad \text{for } h \in H, k \in K$$

(we use the symbol $\bullet$ to emphasize the action) so that (1) can be written

$$(h_1k_1)(h_2k_2) = (h_1\, k_1 \bullet h_2)(k_1k_2). \tag{5.2}$$

The action of K on H by conjugation gives a homomorphism φ of K into $\mathrm{Aut}(H)$, so (2) shows that the multiplication in HK depends only on the multiplication in H, the multiplication in K and the homomorphism φ, hence is defined intrinsically in terms of H and K.

We now use this interpretation to define a group given two groups H and K and a homomorphism φ from K to $\mathrm{Aut}(H)$ (which will turn out to define conjugation in the resulting group).

Theorem 10. Let H and K be groups and let φ be a homomorphism from K into $\mathrm{Aut}(H)$. Let $\bullet$ denote the (left) action of K on H determined by φ. Let G be the set of ordered pairs (h, k) with $h \in H$ and $k \in K$ and define the following multiplication on G:

$$(h_1, k_1)(h_2, k_2) = (h_1\, k_1 \bullet h_2\, ,\, k_1k_2).$$

(1) This multiplication makes G into a group of order $|G| = |H||K|$.

(2) The sets $\{(h, 1) \mid h \in H\}$ and $\{(1, k) \mid k \in K\}$ are subgroups of G and the maps $h \mapsto (h, 1)$ for $h \in H$ and $k \mapsto (1, k)$ for $k \in K$ are isomorphisms of these subgroups with the groups H and K respectively:

$$H \cong \{(h, 1) \mid h \in H\} \quad \text{and} \quad K \cong \{(1, k) \mid k \in K\}.$$

Identifying H and K with their isomorphic copies in G described in (2) we have

(3) $H \trianglelefteq G$

(4) $H \cap K = 1$

(5) for all $h \in H$ and $k \in K$, $khk^{-1} = k \bullet h = \varphi(k)(h)$.

Proof: It is straightforward to check that G is a group under this multiplication using the fact that $\bullet$ is an action of K on H. For example, the associative law is verified

as follows:

$$\begin{aligned}\big((a,x)(b,y)\big)\,(c,z) &= (a\ x{\bullet}b\,,\ xy)(c,z)\\ &= (a\ x{\bullet}b\ (xy){\bullet}c\,,\ xyz)\\ &= (a\ x{\bullet}b\ x{\bullet}(y{\bullet}c)\,,\ xyz)\\ &= (a\ x{\bullet}(b\ y{\bullet}c)\,,\ xyz)\\ &= (a,x)(b\ y{\bullet}c\,,\ yz)\\ &= (a,x)\ \big((b,y)(c,z)\big)\end{aligned}$$

for all $(a,x),(b,y),(c,z) \in G$. We leave as an exercise the verification that (1,1) is the identity of G and that

$$(h,k)^{-1} = (k^{-1}{\bullet}h^{-1}\,,\ k^{-1})$$

for each $(h,k) \in G$. The order of the group G is clearly the product of the orders of H and K, which proves (1).

Let $\widetilde{H} = \{(h,1) \mid h \in H\}$ and $\widetilde{K} = \{(1,k) \mid k \in K\}$. We have

$$(a,1)(b,1) = (a\ 1{\bullet}b\,,\ 1) = (ab,1)$$

for all $a,b \in H$ and

$$(1,x)(1,y) = (1,xy)$$

for all $x,y \in K$, which show that $\widetilde{H}$ and $\widetilde{K}$ are subgroups of G and that the maps in (2) are isomorphisms.

It is clear that $\widetilde{H} \cap \widetilde{K} = 1$, which is (4). Now,

$$\begin{aligned}(1,k)(h,1)(1,k)^{-1} &= \big((1,k)(h,1)\big)(1,k^{-1})\\ &= (k{\bullet}h\,,\ k)(1,k^{-1})\\ &= (k{\bullet}h\ k{\bullet}1\,,\ kk^{-1})\\ &= (k{\bullet}h,1)\end{aligned}$$

so that identifying $(h,1)$ with h and $(1,k)$ with k by the isomorphisms in (2) we have $khk^{-1} = k{\bullet}h$, which is (5).

Finally, we have just seen that (under the identifications in (2)) $K \le N_G(H)$. Since $G = HK$ and certainly $H \le N_G(H)$, we have $N_G(H) = G$, i.e. $H \trianglelefteq G$, which proves (3) and completes the proof.

Definition. Let H and K be groups and let φ be a homomorphism from K into $\mathrm{Aut}(H)$. The group described in Theorem 10 is called the *semidirect product* of H and K with respect to φ and will be denoted by $H \rtimes_\varphi K$ (when there is no danger of confusion we shall simply write $H \rtimes K$).

The notation is chosen to remind us that the copy of H in $H \rtimes K$ is the normal "factor" and that the construction of a semidirect products is not symmetric in H and K (unlike that of a direct product). Before giving some examples we clarify exactly when the semidirect product of H and K is their direct product (in particular, we see that direct products are a special case of semidirect products). See also Exercise 1.

Proposition 11. Let H and K be groups and let $\varphi : K \to \mathrm{Aut}(H)$ be a homomorphism. Then the following are equivalent:

(1) the identity (set) map between $H \rtimes K$ and $H \times K$ is a group homomorphism (hence an isomorphism)

(2) φ is the trivial homomorphism from K into $\mathrm{Aut}(H)$

(3) $K \trianglelefteq H \rtimes K$.

Proof: (1) $\Rightarrow$ (2) By definition of the group operation in $H \rtimes K$

$$(h_1, k_1)(h_2, k_2) = (h_1\, k_1{\bullet}h_2, k_1k_2)$$

for all $h_1, h_2 \in H$ and $k_1, k_2 \in K$. By assumption (1) however, $(h_1, k_1)(h_2, k_2) = (h_1h_2, k_1k_2)$. Equating the first factors of these ordered pairs gives $k_1{\bullet}h_2 = h_2$ for all $h_2 \in H$ and all $k_1 \in K$, i.e. K acts trivially on H. This is (2).

(2) $\Rightarrow$ (3) If φ is trivial, then the action of K on H is trivial, so that the elements of H commute with those of K by Theorem 10(5). In particular, H normalizes K. Since K normalizes itself, $G = HK$ normalizes K, which is (3).

(3) $\Rightarrow$ (1) If K is normal in $H \rtimes K$ then (as in the proof of Theorem 9) for all $h \in H$ and $k \in K$, $[h, k] \in H \cap K = 1$. Thus $hk = kh$ and the action of K on H is trivial. The multiplication in the semidirect product is then the same as that in the direct product:

$$(h_1, k_1)(h_2, k_2) = (h_1h_2, k_1k_2)$$

for all $h_1, h_2 \in H$ and $k_1, k_2 \in K$. This gives (1) and completes the proof.

Examples

In all examples H and K are groups and φ is a homomorphism from K into Aut (H) with associated action of K on H denoted by a dot. Let $G = H \rtimes K$ and as in Theorem 10 we identify H and K as subgroups of G. We shall use Propositions 4.16 and 4.17 to determine homomorphisms φ for some specific groups H. In each of the following examples the proof that φ is a homomorphism is easy (since K will often be cyclic) so the details are omitted.

(1) Let H be any abelian group (even of infinite order) and let $K = \langle x \rangle \cong Z_2$ be the group of order 2. Define $\varphi : K \to \mathrm{Aut}(H)$ by mapping x to the automorphism of inversion on H so that the associated action is $x{\bullet}h = h^{-1}$, for all $h \in H$. Then G contains the subgroup H of index 2 and

$$xhx^{-1} = h^{-1} \qquad \text{for all } h \in H.$$

Of particular interest is the case when H is cyclic: if $H = Z_n$, one recognizes G as D_{2n} and if $H = \mathbb{Z}$ we denote G by D_∞.

(2) We can generalize the preceding example in a number of ways. One way is to let H be any abelian group and to let $K = \langle x \rangle \cong Z_{2n}$ be cyclic of order $2n$. Define φ again by mapping x to inversion, so that x^2 acts as the identity on H. In G, $xhx^{-1} = h^{-1}$ and $x^2hx^{-2} = h$ for all $h \in H$. Thus $x^2 \in Z(G)$. In particular, if $H = Z_3$ and $K = Z_4$, G is a non-abelian group of order 12 which is not isomorphic to A_4 or D_{12} (since its Sylow 2-subgroup, K, is *cyclic* of order 4).

(3) Following up on the preceding example let $H = \langle h \rangle \cong Z_{2^n}$ and let $K = \langle x \rangle \cong Z_4$ with $xhx^{-1} = h^{-1}$ in G. As noted above, $x^2 \in Z(G)$. Since x inverts h (i.e. inverts

H), x inverts the unique subgroup $\langle z\rangle$ of order 2 in H, where $z = h^{2^{n-1}}$. Thus $xzx^{-1} = z^{-1} = z$, so x centralizes z. It follows that $z \in Z(G)$. Thus $x^2z \in Z(G)$ hence $\langle x^2z\rangle \trianglelefteq G$. Let $\overline{G} = G/\langle x^2z\rangle$. Since x^2 and z are distinct commuting elements of order 2, the order of x^2z is 2, so $|\overline{G}| = \frac{1}{2}|G| = 2^{n+1}$. By factoring out the product x^2z to form $\overline{G}$ we identify x^2 and $h^{2^{n-1}}$ in the quotient. In particular, when $n = 2$, both $\bar{x}$ and $\bar{h}$ have order 4, $\bar{x}$ inverts $\bar{h}$ and $\bar{h}^2 = \bar{x}^2$. It follows that $\overline{G} \cong Q_8$ in this case. In general, one can check that $\overline{G}$ has a unique subgroup of order 2 (namely $\langle\bar{x}^2\rangle$) which equals the center of $\overline{G}$. The group $\overline{G}$ is called the *generalized quaternion group* of order 2^{n+1} and is denoted by $Q_{2^{n+1}}$:

$$Q_{2^{n+1}} = \langle h, y \mid h^{2^n} = x^4 = 1,\ x^{-1}hx = h^{-1},\ h^{2^{n-1}} = x^2\rangle.$$

(4) Let $H = \mathbb{Q}$ (under addition) and let $K = \langle x\rangle \cong \mathbb{Z}$. Define φ by mapping x to the map "multiplication by 2" on H, so that x acts on $h \in H$ by $x{\bullet}h = 2h$. Note that multiplication by 2 is an automorphism of H because it has a 2-sided inverse, namely multiplication by $\frac{1}{2}$. In the group G, $\mathbb{Z} \leq \mathbb{Q}$ and the conjugate $x\mathbb{Z}x^{-1}$ of $\mathbb{Z}$ is a *proper* subgroup of $\mathbb{Z}$ (namely $2\mathbb{Z}$). Thus $x \notin N_G(\mathbb{Z})$ even though $x\mathbb{Z}x^{-1} \leq \mathbb{Z}$ (note that $x^{-1}\mathbb{Z}x$ is not contained in $\mathbb{Z}$). This shows that in order to prove an element g normalizes a subgroup A in an *infinite* group it is not sufficient in general to show that the conjugate of A by g is just *contained* in A (which is sufficient for finite groups).

(5) For H any group let $K = \mathrm{Aut}(H)$ with φ the identity map from K to $\mathrm{Aut}(H)$. The semidirect product $H \rtimes \mathrm{Aut}(H)$ is called the *holomorph* of H and will be denoted by $\mathrm{Hol}(H)$. Some holomorphs are described below; verifications of these isomorphisms are given as exercises at the end of this chapter.

(a) $\mathrm{Hol}(Z_2 \times Z_2) \cong S_4$.

(b) If $|G| = n$ and $\pi : G \to S_n$ is the left regular representation (Section 4.2), then $N_{S_n}(\pi(G)) \cong \mathrm{Hol}(G)$. In particular, since the left regular representation of a generator of Z_n is an n-cycle in S_n we obtain that for any n-cycle $(1\,2\,\ldots\,n)$:

$$N_{S_n}(\langle(1\,2\,\ldots\,n)\rangle) \cong \mathrm{Hol}(Z_n) = Z_n \rtimes \mathrm{Aut}(Z_n).$$

Note that the latter group has order $n\varphi(n)$.

(6) Let p and q be primes with $p < q$, let $H = Z_q$ and let $K = Z_p$. We have already seen that if p does not divide $q - 1$ then every group of order pq is cyclic (see the example following Proposition 4.16). This is consistent with the fact that if p does not divide $q - 1$, there is no nontrivial homomorphism from Z_p into $\mathrm{Aut}(Z_q)$ (the latter group is cyclic of order $q - 1$ by Proposition 4.17). Assume now that $p \mid q - 1$. By Cauchy's Theorem, $\mathrm{Aut}(Z_q)$ contains a subgroup of order p (which is unique because $\mathrm{Aut}(Z_q)$ is cyclic). Thus there is a nontrivial homomorphism, φ, from K into $\mathrm{Aut}(H)$. The associated group $G = H \rtimes K$ has order pq and K is not normal in G (Proposition 11). In particular, G is non-abelian. We shall prove shortly that G is (up to isomorphism) the unique non-abelian group of order pq. If $p = 2$, G must be isomorphic to D_{2q}.

(7) Let p be an odd prime. We construct two nonisomorphic non-abelian groups of order p^3 (we shall later prove that any non-abelian group of order p^3 is isomorphic to one of these two).

Let $H = Z_p \times Z_p$ and let $K = Z_p$. By Proposition 4.17, $\mathrm{Aut}(H) \cong GL_2(\mathbb{F}_p)$ and $|GL_2(\mathbb{F}_p)| = (p^2 - 1)(p^2 - p)$. Since $p \mid |\mathrm{Aut}(H)|$, by Cauchy's Theorem H has an automorphism of order p. Thus there is a nontrivial homomorphism, φ, from K into $\mathrm{Aut}(H)$ and so the associated group $H \rtimes K$ is a non-abelian group of order

p^3. More explicitly, if $H = \langle a \rangle \times \langle b \rangle$, and x is a generator for K then x acts on a and b by

$$x \bullet a = ab \quad \text{and} \quad x \bullet b = b$$

which defines the action of x on all of H. With respect to the $\mathbb{F}_p$-basis a, b of the 2-dimensional vector space H the action of x (which can be considered in additive notation as a nonsingular linear transformation) has matrix

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in GL_2(\mathbb{F}_p).$$

The resulting semidirect product has presentation:

$$\langle x, a, b \mid x^p = a^p = b^p = 1,\ ab = ba,\ xax^{-1} = ab,\ xbx^{-1} = b \rangle$$

(in fact, this group is generated by $\{x, a\}$).

Next let $H = Z_{p^2}$ and $K = Z_p$. Again by Proposition 4.17, $\mathrm{Aut}(H) \cong Z_{p(p-1)}$, so H admits an automorphism of order p. Thus there is a nontrivial homomorphism, φ, from K into $\mathrm{Aut}(H)$ and so the group $H \rtimes K$ is non-abelian and of order p^3. More explicitly, if $H = \langle y \rangle$, and x is a generator for K then x acts on y by

$$x \bullet y = y^{1+p}.$$

The resulting semidirect product has presentation:

$$\langle x, y \mid x^p = y^{p^2} = 1,\ xyx^{-1} = y^{1+p} \rangle.$$

The exercises outline a proof that these two groups are not isomorphic (the former contains no element of order p^2 and the latter clearly does, namely y).

(8) Let $H = Q_8 \times (Z_2 \times Z_2) = \langle i, j \rangle \times (\langle a \rangle \times \langle b \rangle)$ and let $K = \langle y \rangle \cong Z_3$. The map defined by

$$i \mapsto j \qquad j \mapsto k = ij \qquad a \mapsto b \qquad b \mapsto ab$$

is easily seen to give an automorphism of H of order 3. Let φ be the homomorphism from K to $\mathrm{Aut}(H)$ defined by mapping y to this automorphism, and let G be the associated semidirect product, so that $y \in G$ acts by

$$y \bullet i = j \qquad y \bullet j = k \qquad y \bullet a = b \qquad y \bullet b = ab.$$

The group $G = H \rtimes K$ is a non-abelian group of order 96 with the property that the element $i^2 a \in G'$ but $i^2 a$ cannot be expressed as a single commutator $[x, y]$, for any $x, y \in G$ (checking the latter assertion is an elementary calculation).

As in the case of direct products we now prove a recognition theorem for semidirect products. This theorem will enable us to "break down" or "factor" all groups of certain orders and, as a result, classify groups of those orders. The strategy is discussed in greater detail following this theorem.

Theorem 12. Suppose G is a group with subgroups H and K such that

(1) $H \trianglelefteq G$, and

(2) $H \cap K = 1$.

Let $\varphi : K \to \mathrm{Aut}(H)$ be the homomorphism defined by mapping $k \in K$ to the automorphism of left conjugation by k on H. Then $HK \cong H \rtimes K$. In particular, if

$G = HK$ with H and K satisfying (1) and (2), then G is the semidirect product of H and K.

Proof: Note that since $H \trianglelefteq G$, HK is a subgroup of G. By Proposition 8 every element of HK can be written uniquely in the form hk, for some $h \in H$ and $k \in K$. Thus the map $hk \mapsto (h, k)$ is a *set* bijection from HK onto $H \rtimes K$. The fact that this map is a homomorphism is the computation at the beginning of this section which led us to the formulation of the definition of the semidirect product.

Definition. Let H be a subgroup of the group G. A subgroup K of G is called a *complement* for H in G if $G = HK$ and $H \cap K = 1$.

With this terminology, the criterion for recognizing a semidirect product is simply that there must exist a complement for some proper *normal* subgroup of G. Not every group is the semidirect product of two of its proper subgroups (for example, if the group is simple), but as we have seen, the notion of a semidirect product greatly increases our list of known groups.

Some Classifications

We now apply Theorem 12 to classify groups of order n for certain values of n. The basic idea in each of the following arguments is to

(**a**) show every group of order n has proper subgroups H and K satisfying the hypothesis of Theorem 12 with $G = HK$
(**b**) find all possible isomorphism types for H and K
(**c**) for each pair H, K found in (b) find all possible homomorphisms $\varphi : K \to \mathrm{Aut}(H)$
(**d**) for each triple H, K, φ found in (c) form the semidirect product $H \rtimes K$ (so any group G of order n is isomorphic to one of these explicitly constructed groups) and among all these semidirect products determine which pairs are isomorphic. This results in a list of the distinct isomorphism types of groups of order n.

In order to start this process we must first find subgroups H and K (of an arbitrary group G of order n) satisfying the above conditions. In the case of "small" values of n we can often do this by Sylow's Theorem. To show *normality* of H we use the conjugacy part of Sylow's Theorem or other normality criteria established in Chapter 4 (e.g. Corollary 4.5). Some of this work has already been done in the examples in Section 4.5. In many of the examples that follow, $|H|$ and $|K|$ are relatively prime, so $H \cap K = 1$ holds by Lagrange's Theorem.

Since H and K are proper subgroups of G one should think of the determination of H and K as being achieved inductively. In the examples we discuss, H and K will have sufficiently small order that we shall know all possible isomorphism types from previous results. For example, in most instances H and K will be of prime or prime squared order.

There will be relatively few possible homomorphisms $\varphi : K \to \mathrm{Aut}(H)$ in our examples, particularly after we take into account certain symmetries (such as replacing one generator of K by another when K is cyclic).

Finally, the semidirect products which emerge from this process will, in our examples, be small in number and we shall find that, for the most part, they are (pair-

wise) *not* isomorphic. In general, this can be a more delicate problem, as Exercise 4 indicates.

We emphasize that this approach to "factoring" every group of some given order n as a semidirect product does not work for arbitrary n. For example, Q_8 is not a semidirect product since no proper subgroup has a complement (although we saw that it is a *quotient* of a semidirect product). Empirically, this process generally works well when the group order n is not divisible by a large power of any prime. At the other extreme, only a small percentage of the groups of order p^α for large α (p a prime) are nontrivial semidirect products.

Example: (Groups of Order *pq*, *p*, and *q* primes with $p < q$)

Let G be any group of order pq, let $P \in Syl_p(G)$ and let $Q \in Syl_q(G)$. In Example 1 of the applications of Sylow's Theorems we proved that $G \cong Q \rtimes P$, for some $\varphi : P \to \mathrm{Aut}(Q)$. Since P and Q are of prime order, they are cyclic. The group $\mathrm{Aut}(Q)$ is cyclic of order $q-1$. If p does not divide $q-1$, the only homomorphism from P to $\mathrm{Aut}(Q)$ is the trivial homomorphism, hence the only semidirect product in this case is the direct product, i.e. G is cyclic.

Consider now the case when $p \mid q-1$ and let $P = \langle y \rangle$. Since $\mathrm{Aut}(Q)$ is cyclic it contains a unique subgroup of order p, say $\langle \gamma \rangle$, and any homomorphism $\varphi : P \to \mathrm{Aut}(Q)$ must map y to a power of γ. There are therefore p homomorphisms $\varphi_i : P \to \mathrm{Aut}(Q)$ given by $\varphi_i(y) = \gamma^i$, $0 \le i \le p-1$. Since φ_0 is the trivial homomorphism, $Q \rtimes_{\varphi_0} P \cong Q \times P$ as before. Each φ_i for $i \neq 0$ gives rise to a non-abelian group, G_i, of order pq. It is straightforward to check that these groups are all isomorphic because for each φ_i, $i > 0$, there is some generator y_i of P such that $\varphi_i(y_i) = \gamma$. Thus, up to a choice for the (arbitrary) generator of P, these semidirect products are all the same (see Exercise 6. See also Exercise 28 of Section 4.3).

Example: (Groups of Order 30)

By the examples following Sylow's Theorem every group G of order 30 contains a subgroup H of order 15. By the preceding example H is cyclic and H is normal in G (index 2). By Sylow's Theorem there is a subgroup K of G of order 2. Thus $G = HK$ and $H \cap K = 1$ so $G \cong H \rtimes K$, for some $\varphi : K \to \mathrm{Aut}(H)$. By Proposition 4.16,

$$\mathrm{Aut}(Z_{15}) \cong (\mathbb{Z}/15\mathbb{Z})^\times \cong Z_4 \times Z_2.$$

The latter isomorphism can be computed directly, or one can use Exercise 11 of the preceding section: writing H as $\langle a \rangle \times \langle b \rangle \cong Z_5 \times Z_3$, we have (since these two subgroups are characteristic in H)

$$\mathrm{Aut}(H) \cong \mathrm{Aut}(Z_5) \times \mathrm{Aut}(Z_3).$$

In particular, $\mathrm{Aut}(H)$ contains precisely three elements of order 2, whose actions on $\langle a \rangle \times \langle b \rangle$ are the following:

$$\left\{ \begin{matrix} a & \mapsto & a \\ b & \mapsto & b^{-1} \end{matrix} \right\} \qquad \left\{ \begin{matrix} a & \mapsto & a^{-1} \\ b & \mapsto & b \end{matrix} \right\} \qquad \left\{ \begin{matrix} a & \mapsto & a^{-1} \\ b & \mapsto & b^{-1} \end{matrix} \right\}.$$

Thus there are three nontrivial homomorphisms from K into $\mathrm{Aut}(H)$ given by sending the generator of K into one of these three elements of order 2 (as usual, the trivial homomorphism gives the direct product: $H \times K \cong Z_{30}$).

Let $K = \langle k \rangle$. If the homomorphism $\varphi_1 : K \to \mathrm{Aut}(H)$ is defined by mapping k to the first automorphism above (so that $k \bullet a = a$ and $k \bullet b = b^{-1}$ gives the action of k

on H) then $G_1 = H \rtimes_{\varphi_1} K$ is easily seen to be isomorphic to $Z_5 \times D_6$ (note that in this semidirect product k centralizes the element a of H of order 5, so the factorization as a direct product is $\langle a \rangle \times \langle b, k \rangle$).

If φ_2 is defined by mapping k to the second automorphism above, then $G_2 = H \rtimes_{\varphi_2} K$ is easily seen to be isomorphic to $Z_3 \times D_{10}$ (note that in this semidirect product k centralizes the element b of H of order 3, so the factorization as a direct product is $\langle b \rangle \times \langle a, k \rangle$).

If φ_3 is defined by mapping k to the third automorphism above then $G_3 = H \rtimes_{\varphi_3} K$ is easily seen to be isomorphic to D_{30}.

Note that these groups are all nonisomorphic since their centers have orders 30 (in the abelian case), 5 (for G_1), 3 (for G_2), and 1 (for G_3).

We emphasize that although (in hindsight) this procedure does not give rise to any groups we could not already have constructed using only direct products, the argument proves that this is the *complete* list of isomorphism types of groups of order 30.

Example: (Groups of Order 12)

Let G be a group of order 12, let $V \in Syl_2(G)$ and let $T \in Syl_3(G)$. By the discussion of groups of order 12 in Section 4.5 we know that either V or T is normal in G (for purposes of illustration we shall not invoke the full force of our results from Chapter 4, namely that either $T \trianglelefteq G$ or $G \cong A_4$). By Lagrange's Theorem $V \cap T = 1$. Thus G is a semidirect product. Note that $V \cong Z_4$ or $Z_2 \times Z_2$ and $T \cong Z_3$.

Case 1: $V \trianglelefteq G$

We must determine all possible homomorphisms from T into $\mathrm{Aut}(V)$. If $V \cong Z_4$, then $\mathrm{Aut}(V) \cong Z_2$ and there are no nontrivial homomorphisms from T into $\mathrm{Aut}(V)$. Thus the only group of order 12 with a normal cyclic Sylow 2-subgroup is Z_{12}.

Assume therefore that $V \cong Z_2 \times Z_2$. In this case $\mathrm{Aut}(V) \cong S_3$ and there is a unique subgroup of $\mathrm{Aut}(V)$ of order 3, say $\langle \gamma \rangle$. Thus if $T = \langle y \rangle$, there are three possible homomorphisms from T into $\mathrm{Aut}(V)$:

$$\varphi_i : T \to \mathrm{Aut}(V) \quad \text{defined by} \quad \varphi_i(y) = \gamma^i, \quad i = 0, 1, 2.$$

As usual, φ_0 is the trivial homomorphism which gives rise to the direct product: $Z_2 \times Z_2 \times Z_3$. Homomorphisms φ_1 and φ_2 give rise to isomorphic semidirect products because they differ only in the choice of a generator for T (i.e. $\varphi_1(y) = \gamma$ and $\varphi_2(y') = \gamma$, where $y' = y^2$ and y' is another choice of generator for T – see also Exercise 6). The unique non-abelian group in this case is A_4.

Case 2: $T \trianglelefteq G$

We must determine all possible homomorphisms from V into $\mathrm{Aut}(T)$. Note that $\mathrm{Aut}(T) = \langle \lambda \rangle \cong Z_2$, where λ inverts T. If $V = \langle x \rangle \cong Z_4$, there are precisely two homomorphisms from V into $\mathrm{Aut}(T)$, the trivial homomorphism and the homomorphism which sends x to λ. As usual, the trivial homomorphism gives rise to the direct product: $Z_3 \times Z_4 \cong Z_{12}$. The nontrivial homomorphism gives the semidirect product which was discussed in Example 2 following Proposition 11 of this section.

Finally, assume $V = \langle a \rangle \times \langle b \rangle \cong Z_2 \times Z_2$. There are precisely three nontrivial homomorphisms from V into $\mathrm{Aut}(T)$ determined by specifying their kernels as one of the three subgroups of order 2 in V. For example, $\varphi_1(a) = \lambda$ and $\varphi_1(b) = \lambda$ has kernel $\langle ab \rangle$, that is, in this semidirect product both a and b act by inverting T and ab centralizes T. If φ_2 and φ_3 have kernels $\langle a \rangle$ and $\langle b \rangle$, respectively, then one easily checks that the

resulting three semidirect products are all isomorphic to $S_3 \times Z_2$, where the Z_2 direct factor is the kernel of φ_i. For example,

$$T \rtimes_{\varphi_1} V = \langle a, T \rangle \times \langle ab \rangle.$$

In summary, there are precisely 5 groups of order 12, three of which are non-abelian.

Example: (Groups of Order p^3, p an odd prime)

Let G be a group of order p^3, p an odd prime, and assume G is not cyclic. By Exercise 9 of the previous section the map $x \mapsto x^p$ is a homomorphism from G into $Z(G)$. Since $|Z(G)| = p$, the kernel of this homomorphism has order p^2 or p^3. In the former case G must contain an element of order p^2 and in the latter case every nonidentity element of G has order p.

Case 1: G has an element of order p^2

Let x be an element of order p^2 and let $H = \langle x \rangle$. Note that since H has index p, H is normal in G by Corollary 4.5. If E is the kernel of the pth power map, then in this case $E \cong Z_p \times Z_p$ and $E \cap H = \langle x^p \rangle$. Let y be any element of $E - H$ and let $K = \langle y \rangle$. By construction, $H \cap K = 1$ and so G is isomorphic to $Z_{p^2} \rtimes Z_p$, for some $\varphi : K \to \mathrm{Aut}(H)$. If φ is the trivial homomorphism, $G \cong Z_{p^2} \times Z_p$, so we need only consider the nontrivial homomorphisms. By Proposition 4.17 $\mathrm{Aut}(H) \cong Z_{p(p-1)}$ is cyclic and so contains a unique subgroup of order p, explicitly given by $\langle \gamma \rangle$ where

$$\gamma(x) = x^{1+p}.$$

As usual, up to choice of a generator for the cyclic group K, there is only one nontrivial homomorphism, φ, from K into $\mathrm{Aut}(H)$, given by $\varphi(y) = \gamma$; hence up to isomorphism there is a unique non-abelian group $H \rtimes K$ in this case. This group is described in Example 7 above.

Case 2: every nonidentity element of G has order p

In this case let H be any subgroup of G of order p^2 (see Exercise 29, Section 4.3). Necessarily $H \cong Z_p \times Z_p$. Let $K = \langle y \rangle$ for any element y of $G - H$. Since H has index p, $H \trianglelefteq G$ and since K has order p but is not contained in H, $H \cap K = 1$. Then G is isomorphic to $(Z_p \times Z_p) \rtimes Z_p$, for some $\varphi : K \to \mathrm{Aut}(H)$. If φ is trivial, $G \cong Z_p \times Z_p \times Z_p$ (the elementary abelian group), so we may assume φ is nontrivial. By Proposition 4.17,

$$\mathrm{Aut}(H) \cong GL_2(\mathbb{F}_p)$$

so $|\mathrm{Aut}(H)| = (p^2-1)(p^2-p)$. Note that a Sylow p-subgroup of $\mathrm{Aut}(H)$ has order p so all subgroups of order p in $\mathrm{Aut}(H)$ are conjugate in $\mathrm{Aut}(H)$ by Sylow's Theorem. Explicitly, (as discussed in Example 7 above) every subgroup of order p in $\mathrm{Aut}(H)$ is conjugate to $\langle \gamma \rangle$, where if $H = \langle a \rangle \times \langle b \rangle$, the automorphism γ is defined by

$$\gamma(a) = ab \quad \text{and} \quad \gamma(b) = b.$$

With respect to the $\mathbb{F}_p$–basis a, b of the 2-dimensional vector space H the automorphism has matrix

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \in GL_2(\mathbb{F}_p).$$

Thus (again quoting Exercise 6) there is a unique isomorphism type of semidirect product in this case. Finally, since the two non-abelian groups have different orders for the kernels of the pth power maps, they are not isomorphic. A presentation for this group is also given in Example 7 above.

EXERCISES

Let H and K be groups, let φ be a homomorphism from K into $\mathrm{Aut}(H)$ and, as usual, identify H and K as subgroups of $G = H \rtimes_\varphi K$.

1. Prove that $C_K(H) = \ker \varphi$ (recall that $C_K(H) = C_G(H) \cap K$).
2. Prove that $C_H(K) = N_H(K)$.
3. In Example 1 following the proof of Proposition 11 prove that every element of $G - H$ has order 2. Prove that G is abelian if and only if $h^2 = 1$ for all $h \in H$.
4. Let $p = 2$ and check that the construction of the two non-abelian groups of order p^3 is valid in this case. Prove that *both* resulting groups are isomorphic to D_8.
5. Let $G = \mathrm{Hol}(Z_2 \times Z_2)$.
   (a) Prove that $G = H \rtimes K$ where $H = Z_2 \times Z_2$ and $K \cong S_3$. Deduce that $|G| = 24$.
   (b) Prove that G is isomorphic to S_4. [Obtain a homomorphism from G into S_4 by letting G act on the left cosets of K. Use Exercise 1 to show this representation is faithful.]
6. Assume that K is a cyclic group, H is an arbitrary group and φ_1 and φ_2 are homomorphisms from K into $\mathrm{Aut}(H)$ such that $\varphi_1(K)$ and $\varphi_2(K)$ are conjugate subgroups of $\mathrm{Aut}(H)$. Prove by constructing an explicit isomorphism that $H \rtimes_{\varphi_1} K \cong H \rtimes_{\varphi_2} K$ (in particular, if the subgroups $\varphi_1(K)$ and $\varphi_2(K)$ are equal in $\mathrm{Aut}(H)$, then the resulting semidirect products are isomorphic). [Suppose $\sigma\varphi_1(K)\sigma^{-1} = \varphi_2(K)$ so that for some $a \in \mathbb{Z}$ we have $\sigma\varphi_1(k)\sigma^{-1} = \varphi_2(k)^a$ for all $k \in K$. Show that the map $\psi : H \rtimes_{\varphi_1} K \to H \rtimes_{\varphi_2} K$ defined by $\psi((h, k)) = (\sigma(h), k^a)$ is a homomorphism. Show ψ is bijective by constructing a 2-sided inverse.]
7. This exercise describes all the groups of order 56 (there are thirteen isomorphism types).
   (a) Prove that there are three abelian groups of order 56.
   (b) Prove that every group of order 56 has either a normal Sylow 2-subgroup or a normal Sylow 7-subgroup.
   (c) Construct the following non-abelian groups of order 56 which have a normal Sylow 7-subgroup and whose Sylow 2-subgroup S is as specified:
   one group when $S \cong Z_2 \times Z_2 \times Z_2$
   two nonisomorphic groups when $S \cong Z_4 \times Z_2$
   one group when $S \cong Z_8$
   two nonisomorphic groups when $S \cong Q_8$
   three nonisomorphic groups when $S \cong D_8$.
   [For a particular S, two groups are not isomorphic if the kernels of the maps from S into $\mathrm{Aut}(Z_7)$ are not isomorphic.]
   (d) Prove that if G is a non-abelian group of order 56 with a normal Sylow 2-subgroup S, then $S \cong Z_2 \times Z_2 \times Z_2$. [Let an element of order 7 act by conjugation on the seven nonidentity elements of S and deduce that they all have the same order.]
   (e) Prove that there is a unique non-abelian group of order 56 with a normal Sylow 2-subgroup. [To prove existence use the fact that $|GL_3(\mathbb{F}_2)| = 168$. For uniqueness use Exercise 6.]

   It is not too difficult to show that every group of order 56 is isomorphic to one of the thirteen groups above.

8. Construct a non-abelian group of order 75. Classify all groups of order 75 (there are three of them). [Use Exercise 6 to show that the non-abelian group is unique.] (The classification of groups of order pq^2, where p and q are primes with $p < q$ and p not dividing $q - 1$, is quite similar.)

9. Show that the following matrix is an element of order 5 in $GL_2(\mathbb{F}_{19})$:

$$\begin{pmatrix} 0 & -1 \\ 1 & 4 \end{pmatrix}.$$

Use this matrix to construct a non-abelian group of order 1805 and give a presentation of this group. Classify groups of order 1805 (there are three isomorphism types). [Use Exercise 6 to prove uniqueness of the non-abelian group.] (A general method for finding elements of prime order in $GL_n(\mathbb{F}_p)$ is described in the exercises in Section 12.2; this particular matrix of order 5 in $GL_2(\mathbb{F}_{19})$ appears in Exercise 16 of that section as an illustration of the method.)

10. This exercise classifies the groups of order 147 (there are six isomorphism types).
    (a) Prove that there are two abelian groups of order 147.
    (b) Prove that every group of order 147 has a normal Sylow 7-subgroup.
    (c) Prove that there is a unique non-abelian group whose Sylow 7-subgroup is cyclic.
    (d) Let $t_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $t_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ be elements of $GL_2(\mathbb{F}_7)$. Prove that $P = \langle t_1, t_2 \rangle$ is a Sylow 3-subgroup of $GL_2(\mathbb{F}_7)$ and that $P \cong Z_3 \times Z_3$. Deduce that every subgroup of $GL_2(\mathbb{F}_7)$ of order 3 is conjugate in $GL_2(\mathbb{F}_7)$ to a subgroup of P.
    (e) By Example 3 in Section 1 the group P has four subgroups of order 3 and these are: $P_1 = \langle t_1 \rangle$, $P_2 = \langle t_2 \rangle$, $P_3 = \langle t_1 t_2 \rangle$, and $P_4 = \langle t_1 t_2^2 \rangle$. For $i = 1, 2, 3, 4$ let $G_i = Z_7 \rtimes_{\varphi_i} Z_3$, where φ_i is an isomorphism of Z_3 with the subgroup P_i of $\mathrm{Aut}(Z_7)$. For each i describe G_i in terms of generators and relations. Deduce that $G_1 \cong G_2$.
    (f) Prove that G_1 is not isomorphic to either G_3 or G_4. [Show that the center of G_1 has order 7 whereas the centers of G_3 and G_4 are trivial.]
    (g) Prove that G_3 is not isomorphic to G_4. [Show that every subgroup of order 7 in G_3 is normal in G_3 but that G_4 has subgroups of order 7 that are not normal.]
    (h) Classify the groups of order 147 by showing that the six nonisomorphic groups described above (two from part (a), one from part (c) and G_1, G_3, and G_4) are all the groups of order 147. [Use Exercise 6 and part (d).] (The classification of groups of order pq^2, where p and q are primes with $p < q$ and $p \mid q - 1$, is quite similar.)

11. Classify groups of order 28 (there are four isomorphism types).

12. Classify the groups of order 20 (there are five isomorphism types).

13. Classify groups of order $4p$, where p is an odd prime. [There are four isomorphism types when $p \equiv 3 \pmod 4$ and five isomorphism types when $p \equiv 1 \pmod 4$.]

14. This exercise classifies the groups of order 60 (there are thirteen isomorphism types). Let G be a group of order 60, let P be a Sylow 5-subgroup of G and let Q be a Sylow 3-subgroup of G.
    (a) Prove that if P is not normal in G then $G \cong A_5$. [See Section 4.5.]
    (b) Prove that if $P \trianglelefteq G$ but Q is not normal in G then $G \cong A_4 \times Z_5$. [Show in this case that $P \leq Z(G)$, $G/P \cong A_4$, a Sylow 2-subgroup T of G is normal and $TQ \cong A_4$.]
    (c) Prove that if both P and Q are normal in G then $G \cong Z_{15} \rtimes T$ where $T \cong Z_4$ or $Z_2 \times Z_2$. Show in this case that there are six isomorphism types when T is cyclic (one abelian) and there are five isomorphism types when T is the Klein 4-group (one abelian). [Use the same ideas as in the classifications of groups of orders 30 and 20.]

15. Let p be an odd prime. Prove that every element of order 2 in $GL_2(\mathbb{F}_p)$ is conjugate to a diagonal matrix with ± 1's on the diagonal. Classify the groups of order $2p^2$. [If A is a 2×2 matrix with $A^2 = I$ and v_1, v_2 is a basis for the underlying vector space, look at A acting on the vectors $w_1 = v_1 + v_2$ and $w_2 = v_1 - v_2$.]

16. Show that there are exactly 4 distinct homomorphisms from Z_2 into $\mathrm{Aut}(Z_8)$. Prove that the resulting semidirect products are the groups: $Z_8 \times Z_2$, D_{16}, the quasidihedral group QD_{16} and the modular group M (cf. the exercises in Section 2.5).

17. Show that for any $n \geq 3$ there are exactly 4 distinct homomorphisms from Z_2 into $\mathrm{Aut}(Z_{2^n})$. Prove that the resulting semidirect products give 4 nonisomorphic groups of order 2^{n+1}. [Recall Exercises 21 to 23 in Section 2.3.] (These four groups together with the cyclic group and the generalized quaternion group, $Q_{2^{n+1}}$, are all the groups of order 2^{n+1} which possess a cyclic subgroup of index 2.)

18. Show that if H is any group then there is a group G that contains H as a normal subgroup with the property that for every automorphism σ of H there is an element $g \in G$ such that conjugation by g when restricted to H is the given automorphism σ, i.e. every automorphism of H is obtained as an inner automorphism of G restricted to H.

19. Let H be a group of order n, let $K = \mathrm{Aut}(H)$ and form $G = \mathrm{Hol}(H) = H \rtimes K$ (where φ is the identity homomorphism). Let G act by left multiplication on the left cosets of K in G and let π be the associated permutation representation $\pi : G \to S_n$.
    (a) Prove that the elements of H are coset representatives for the left cosets of K in G and with this choice of coset representatives π restricted to H is the regular representation of H.
    (b) Prove that $\pi(G)$ equals the normalizer in S_n of $\pi(H)$. Deduce that under the regular representation of any finite group H of order n, the normalizer in S_n of the image of H is isomorphic to $\mathrm{Hol}(H)$. [Show $|G| = |N_{S_n}(\pi(H))|$ using Exercises 1 and 2 above.]
    (c) Deduce that the normalizer of the group generated by an n-cycle in S_n is isomorphic to $\mathrm{Hol}(Z_n)$ and has order $n\varphi(n)$.

20. Let p be an odd prime. Prove that if P is a non-cyclic p-group then P contains a normal subgroup U with $U \cong Z_p \times Z_p$. Deduce that for odd primes p a p-group that contains a unique subgroup of order p is cyclic. (For $p = 2$ it is a theorem that the generalized quaternion groups Q_{2^n} are the only non-cyclic 2-groups which contain a unique subgroup of order 2). [Proceed by induction on $|P|$. Let Z be a subgroup of order p in $Z(P)$ and let $\overline{P} = P/Z$. If $\overline{P}$ is cyclic then P is abelian by Exercise 36 in Section 3.1 — show the result is true for abelian groups. When $\overline{P}$ is not cyclic use induction to produce a normal subgroup $\overline{H}$ of $\overline{P}$ with $\overline{H} \cong Z_p \times Z_p$. Let H be the complete preimage of $\overline{H}$ in P, so $|H| = p^3$. Let $H_0 = \{x \in H \mid x^p = 1\}$ so that H_0 is a characteristic subgroup of H of order p^2 or p^3 by Exercise 9 in Section 4. Show that a suitable subgroup of H_0 gives the desired normal subgroup U.]

21. Let p be an odd prime and let P be a p-group. Prove that if every subgroup of P is normal then P is abelian. (Note that Q_8 is a non-abelian 2-group with this property, so the result is false for $p = 2$.) [Use the preceding exercises and Exercise 15 of Section 4.]

22. Let F be a field let n be a positive integer and let G be the group of upper triangular matrices in $GL_n(F)$ (cf. Exercise 16, Section 2.1).
    (a) Prove that G is the semidirect product $U \rtimes D$ where U is the set of upper triangular matrices with 1's down the diagonal (cf. Exercise 17, Section 2.1) and D is the set of diagonal matrices in $GL_n(F)$.
    (b) Let $n = 2$. Recall that $U \cong F$ and $D \cong F^\times \times F^\times$ (cf. Exercise 11 in Section 3.1). Describe the homomorphism from D into $\mathrm{Aut}(U)$ explicitly in terms of these isomorphisms (i.e. show how each element of $F^\times \times F^\times$ acts as an automorphism on F).

23. Let K and L be groups, let n be a positive integer, let $\rho : K \to S_n$ be a homomorphism and let H be the direct product of n copies of L. In Exercise 8 of Section 1 an injective homomorphism ψ from S_n into $\mathrm{Aut}(H)$ was constructed by letting the elements of S_n permute the n factors of H. The composition $\psi \circ \rho$ is a homomorphism from G into $\mathrm{Aut}(H)$. The *wreath product* of L by K is the semidirect product $H \rtimes K$ with respect to this homomorphism, and is denoted by $L \wr K$ (this wreath product depends on the choice of permutation representation ρ of K — if none is given explicitly, ρ is assumed to be the left regular representation of K).

(a) Assume K and L are finite groups and ρ is the left regular representation of K. Find $|L \wr K|$ in terms of $|K|$ and $|L|$.

(b) Let p be a prime, let $K = L = Z_p$ and let ρ be the left regular representation of K. Prove that $Z_p \wr Z_p$ is a non-abelian group of order p^{p+1} and is isomorphic to a Sylow p-subgroup of S_{p^2}. [The p copies of Z_p whose direct product makes up H may be represented by p disjoint p-cycles; these are cyclically permuted by K.]

24. Let n be an integer > 1. Prove the following classification: every group of order n is abelian if and only if $n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_r^{\alpha_r}$, where $p_1, \dots, p_r$ are distinct primes, $\alpha_i = 1$ or 2 for all $i \in \{1, \dots, r\}$ and p_i does not divide $p_j^2 - 1$ for all i and j. [See Exercise 56 in Section 4.5.]

6

Further Topics in Group Theory

6.1 p-GROUPS, NILPOTENT GROUPS AND SOLVABLE GROUPS

Let p be a prime and let G be a finite group of order $p^a n$, where p does not divide n. Recall that a (finite) p-group is any group whose order is a power of p. Sylow's Theorem shows that p-groups abound as subgroups of G and in order to exploit this phenomenon to unravel the structure of finite groups it will be necessary to establish some basic properties of p-groups. In the next section we shall apply these results in many specific instances.

Before giving the results on p-groups we first recall a definition that has appeared in some earlier exercises.

Definition. A *maximal subgroup* of a group G is a proper subgroup M of G such that there are no subgroups H of G with $M < H < G$.

By order considerations every proper subgroup of a finite group is contained in some maximal subgroup. In contrast, infinite groups may or may not have maximal subgroups. For example, $p\mathbb{Z}$ is a maximal subgroup of $\mathbb{Z}$ whereas $\mathbb{Q}$ (under $+$) has no maximal subgroups (cf. Exercise 16 at the end of this section).

We now collect all the properties of p-groups we shall need into an omnibus theorem:

Theorem 1. Let p be a prime and let P be a group of order p^a, $a \geq 1$. Then

(**1**) The center of P is nontrivial: $Z(P) \neq 1$.

(**2**) If H is a nontrivial normal subgroup of P then H intersects the center nontrivially: $H \cap Z(P) \neq 1$. In particular, every normal subgroup of order p is contained in the center.

(**3**) If H is a normal subgroup of P then H contains a subgroup of order p^b that is normal in P for each divisor p^b of $|H|$. In particular, P has a normal subgroup of order p^b for every $b \in \{0, 1, \ldots, a\}$.

(**4**) If $H < P$ then $H < N_P(H)$ (i.e. every proper subgroup of P is a proper subgroup of its normalizer in P).

(**5**) Every maximal subgroup of P is of index p and is normal in P.

Proof: These results rely ultimately on the class equation and it may be useful for the reader to review Section 4.3.

Part (1) is Theorem 8 of Chapter 4 and is also the special case of part (2) when $H = P$. We therefore begin by proving (2); we shall not quote Theorem 8 of Chapter 4 although the argument that follows is only a slight generalization of the one in Chapter 4. Let H be a nontrivial normal subgroup of P. Recall that for each conjugacy class $\mathcal{C}$ of P, either $\mathcal{C} \subseteq H$ or $\mathcal{C} \cap H = \emptyset$ because H is normal (this easy fact was shown in a remark preceding Theorem 4.12). Pick representatives of the conjugacy classes of P:

$$a_1, a_2, \ldots, a_r$$

with $a_1, \ldots, a_k \in H$ and $a_{k+1}, \ldots, a_r \notin H$. Let $\mathcal{C}_i$ be the conjugacy class of a_i in P, for all i. Thus

$$\mathcal{C}_i \subseteq H, \quad 1 \leq i \leq k \qquad \text{and} \qquad \mathcal{C}_i \cap H = \emptyset, \quad k+1 \leq i \leq r.$$

By renumbering $a_1, \ldots, a_k$ if necessary we may assume $a_1, \ldots, a_s$ represent classes of size 1 (i.e. are in the center of P) and $a_{s+1}, \ldots, a_k$ represent classes of size > 1. Since H is the disjoint union of these we have

$$|H| = |H \cap Z(P)| + \sum_{i=s+1}^{k} \frac{|P|}{|C_P(a_i)|}.$$

Now p divides $|H|$ and p divides each term in the sum $\sum_{i=s+1}^{k} |P : C_P(a_i)|$ so p divides their difference: $|H \cap Z(P)|$. This proves $H \cap Z(P) \neq 1$. If $|H| = p$, since $H \cap Z(P) \neq 1$ we must have $H \leq Z(P)$. This completes the proof of (2).

Next we prove (3) by induction on a. If $a \leq 1$ or $H = 1$, the result is trivial. Assume therefore that $a > 1$ and $H \neq 1$. By part (2), $H \cap Z(P) \neq 1$ so by Cauchy's Theorem $H \cap Z(P)$ contains a (normal) subgroup Z of order p. Use bar notation to denote passage to the quotient group P/Z. This quotient has order p^{a-1} and $\overline{H} \trianglelefteq \overline{P}$. By induction, for every nonnegative integer b such that p^b divides $|\overline{H}|$ there is a subgroup $\overline{K}$ of $\overline{H}$ of order p^b that is normal in $\overline{P}$. If K is the complete preimage of $\overline{K}$ in P then $|K| = p^{b+1}$. The set of all subgroups of H obtained by this process together with the identity subgroup provides a subgroup of H that is normal in P for each divisor of $|H|$. The second assertion of part (3) is the special case $H = P$. This establishes part (3).

We prove (4) also by induction on $|P|$. If P is abelian then all subgroups of P are normal in P and the result is trivial. We may therefore assume $|P| > p$ (in fact, $|P| > p^2$ by Corollary 4.9). Let H be a proper subgroup of P. Since all elements of $Z(P)$ commute with all elements of P, $Z(P)$ normalizes every subgroup of P. By part (1) we have that $Z(P) \neq 1$. If $Z(P)$ is not contained in H, then H is properly contained in $\langle H, Z(P) \rangle$ and the latter subgroup is contained in $N_P(H)$ so (4) holds. We may therefore assume $Z(P) \leq H$. Use bar notation to denote passage to the quotient $P/Z(P)$. Since $\overline{P}$ has smaller order than P by (1), by induction $\overline{H}$ is properly contained in $N_{\overline{P}}(\overline{H})$. It follows directly from the Lattice Isomorphism Theorem that $N_P(H)$ is the complete preimage in P of $N_{\overline{P}}(\overline{H})$, hence we obtain proper containment of H in its normalizer in this case as well. This completes the induction.

To prove (5) let M be a maximal subgroup of P. By definition, $M < P$ so by part (4), $M < N_P(M)$. By definition of maximality we must therefore have $N_P(M) = P$,

i.e. $M \trianglelefteq P$. The Lattice Isomorphism Theorem shows that P/M is a p-group with no proper nontrivial subgroups because M is a maximal subgroup. By part (3), however, P/M has subgroups of every order dividing $|P/M|$. The only possibility is $|P/M| = p$. This proves (5) and completes the proof of the theorem.

Definition

(1) For any (finite or infinite) group G define the following subgroups inductively:

$$Z_0(G) = 1, \qquad Z_1(G) = Z(G)$$

and $Z_{i+1}(G)$ is the subgroup of G containing $Z_i(G)$ such that

$$Z_{i+1}(G)/Z_i(G) = Z(G/Z_i(G))$$

(i.e. $Z_{i+1}(G)$ is the complete preimage in G of the center of $G/Z_i(G)$ under the natural projection). The chain of subgroups

$$Z_0(G) \leq Z_1(G) \leq Z_2(G) \leq \cdots$$

is called the *upper central series of* G. (The term "upper" indicates that $Z_i(G) \leq Z_{i+1}(G)$.)

(2) A group G is called *nilpotent* if $Z_c(G) = G$ for some $c \in \mathbb{Z}$. The smallest such c is called the *nilpotence class* of G.

One of the exercises at the end of this section shows that $Z_i(G)$ is a characteristic (hence normal) subgroup of G for all i. We use this fact freely from now on.

Remarks:

(1) If G is abelian then G is nilpotent (of class 1, provided $|G| > 1$), because $G = Z(G) = Z_1(G)$. One should think of nilpotent groups as lying between abelian and solvable groups in the hierarchy of structure (recall that solvable groups were introduced in Section 3.4; we shall discuss solvable groups further at the end of this section):

$$\textit{cyclic groups} \subset \textit{abelian groups} \subset \textit{nilpotent groups} \subset \textit{solvable groups} \subset \textit{all groups}$$

(all of the above containments are proper, as we shall verify shortly).

(2) For any finite group there must, by order considerations, be an integer n such that

$$Z_n(G) = Z_{n+1}(G) = Z_{n+2}(G) = \cdots.$$

For example, $Z_n(S_3) = 1$ for all $n \in \mathbb{Z}^+$. Once two terms in the upper central series are the same, the chain stabilizes at that point (i.e. all terms thereafter are equal to these two). For example, if $G = Z_2 \times S_3$,

$$Z(G) = Z_1(G) = Z_2(G) = Z_n(G) \quad \text{has order 2 for all } n.$$

By definition, $Z_n(G)$ is a proper subgroup of G for all n for non-nilpotent groups.

(3) For infinite groups G it may happen that all $Z_i(G)$ are proper subgroups of G (so G is not nilpotent) but

$$G = \bigcup_{i=0}^{\infty} Z_i(G).$$

Groups for which this hold are called *hypernilpotent* — they enjoy some (but not all) of the properties of nilpotent groups. While we shall be dealing mainly with finite nilpotent groups, results that do not involve the notion of order, Sylow subgroups etc. also hold for infinite groups. Even for infinite groups one of the main techniques for dealing with nilpotent groups is induction on the nilpotence class.

Proposition 2. Let p be a prime and let P be a group of order p^a. Then P is nilpotent of nilpotence class at most $a - 1$.

Proof: For each $i \geq 0$, $P/Z_i(P)$ is a p-group, so

$$\text{if } |P/Z_i(P)| > 1 \text{ then } Z(P/Z_i(P)) \neq 1$$

by Theorem 1(1). Thus if $Z_i(P) \neq 1$ then $|Z_{i+1}(P)| \geq p|Z_i(P)|$ and so $|Z_{i+1}(P)| \geq p^{i+1}$. In particular, $|Z_a(P)| \geq p^a$, so $P = Z_a(P)$. Thus P is nilpotent of class $\leq a$. The only way P could be of nilpotence class exactly equal to a would be if $|Z_i(P)| = p^i$ for all i. In this case, however, $Z_{a-2}(P)$ would have index p^2 in P, so P/Z_{a-2} would be abelian (by Corollary 4.9). But then $P/Z_{a-2}(P)$ would equal its center and so $Z_{a-1}(P)$ would equal P, a contradiction. This proves that the class of P is $\leq a - 1$.

Examples

Both D_8 and Q_8 are nilpotent of class 2. More generally, D_{2^n} is nilpotent of class $n - 1$. This can be proved inductively by showing that for $n \geq 3$, $|Z(D_{2^n})| = 2$ and $D_{2^n}/Z(D_{2^n}) \cong D_{2^{n-1}}$ (the details are left as an exercise). If n is not a power of 2, D_{2n} is not nilpotent (cf. Exercise 10).

We now give some equivalent (and often more workable) characterizations of nilpotence for *finite* groups:

Theorem 3. Let G be a finite group, let $p_1, p_2, \ldots, p_s$ be the distinct primes dividing its order and let $P_i \in Syl_{p_i}(G)$, $1 \leq i \leq s$. Then the following are equivalent:

(1) G is nilpotent

(2) if $H < G$ then $H < N_G(H)$, i.e. every proper subgroup of G is a proper subgroup of its normalizer in G

(3) $P_i \trianglelefteq G$ for $1 \leq i \leq s$, i.e. every Sylow subgroup is normal in G

(4) $G \cong P_1 \times P_2 \times \cdots \times P_s$.

Proof: The proof that (1) implies (2) is the same argument as for p-groups – the only fact we needed was if G is nilpotent then so is $G/Z(G)$ — so the details are omitted (cf. the exercises).

To show that (2) implies (3) let $P = P_i$ for some i and let $N = N_G(P)$. Since $P \trianglelefteq N$, Corollary 4.20 gives that P is characteristic in N. Since P char $N \trianglelefteq N_G(N)$ we get that $P \trianglelefteq N_G(N)$. This means $N_G(N) \leq N$ and hence $N_G(N) = N$. By (2) we must therefore have $N = G$, which gives (3).

Next we prove (3) implies (4). For any t, $1 \leq t \leq s$ we show inductively that

$$P_1 P_2 \cdots P_t \cong P_1 \times P_2 \times \cdots \times P_t.$$

Note first that each P_i is normal in G so $P_1 \cdots P_t$ is a subgroup of G. Let $H = P_1 \cdots P_{t-1}$ and let $K = P_t$, so by induction $H \cong P_1 \times \cdots \times P_{t-1}$. In particular, $|H| = |P_1| \cdot |P_2| \cdots |P_{t-1}|$. Since $|K| = |P_t|$, the orders of H and K are relatively prime. Lagrange's Theorem implies $H \cap K = 1$. By definition, $P_1 \cdots P_t = HK$, hence Theorem 5.9 gives

$$HK \cong H \times K = (P_1 \times \cdots \times P_{t-1}) \times P_t \cong P_1 \times \cdots \times P_t$$

which completes the induction. Now take $t = s$ to obtain (4).

Finally, to prove (4) implies (1) use Exercise 1 of Section 5.1 to obtain

$$Z(P_1 \times \cdots \times P_s) \cong Z(P_1) \times \cdots \times Z(P_s).$$

By Exercise 14 in Section 5.1,

$$G/Z(G) = (P_1/Z(P_1)) \times \cdots \times (P_s/Z(P_s)).$$

Thus the hypotheses of (4) also hold for $G/Z(G)$. By Theorem 1, if $P_i \neq 1$ then $Z(P_i) \neq 1$, so if $G \neq 1$, $|G/Z(G)| < |G|$. By induction, $G/Z(G)$ is nilpotent, so by Exercise 6, G is nilpotent. This completes the proof.

Note that the first part of the Fundamental Theorem of Finite Abelian Groups (Theorem 5 in Section 5.2) follows immediately from the above theorem (we shall give another proof later as a consequence of the Chinese Remainder Theorem):

Corollary 4. A finite abelian group is the direct product of its Sylow subgroups.

Next we prove a proposition which will be used later to show that the multiplicative group of a finite field is cyclic (without using the Fundamental Theorem of Finite Abelian Groups):

Proposition 5. If G is a finite group such that for all positive integers n dividing its order, G contains at most n elements x satisfying $x^n = 1$, then G is cyclic.

Proof: Let $|G| = p_1^{\alpha_1} \cdots p_s^{\alpha_s}$ and let P_i be a Sylow p_i-subgroup of G for $i = 1, 2, \ldots, s$. Since $p_i^{\alpha_i} \mid |G|$ and the $p_i^{\alpha_i}$ elements of P_i are solutions of $x^{p_i^{\alpha_i}} = 1$, by hypothesis P_i must contain *all* solutions to this equation in G. It follows that P_i is the unique (hence normal) Sylow p_i-subgroup of G. By Theorem 3, G is the direct product of its Sylow subgroups. By Theorem 1, each P_i possesses a normal subgroup M_i of index p_i. Since $|M_i| = p_i^{\alpha_i - 1}$ and G has at most $p_i^{\alpha_i - 1}$ solutions to $x^{p_i^{\alpha_i - 1}} = 1$, by Lagrange's Theorem (Corollary 9, Section 3.2) M contains all elements x of G satisfying $x^{p_i^{\alpha_i - 1}} = 1$. Thus any element of P_i not contained in M_i satisfies $x^{p_i^{\alpha_i}} = 1$

but $x^{p_i^{\alpha_i - 1}} \neq 1$, i.e. x is an element of order $p_i^{\alpha_i}$. This proves P_i is cyclic for all i, so G is the direct product of cyclic groups of relatively prime order, hence is cyclic.

The next proposition is called Frattini's Argument. We shall apply it to give another characterization of finite nilpotent groups. It will also be a valuable tool in the next section.

Proposition 6. (Frattini's Argument) Let G be a finite group, let H be a normal subgroup of G and let P be a Sylow p-subgroup of H. Then $G = HN_G(P)$ and $|G : H|$ divides $|N_G(P)|$.

Proof: By Corollary 3.15, $HN_G(P)$ is a subgroup of G and $HN_G(P) = N_G(P)H$ since H is a normal subgroup of G. Let $g \in G$. Since $P^g \leq H^g = H$, both P and P^g are Sylow p-subgroups of H. By Sylow's Theorem applied in H, there exists $x \in H$ such that $P^g = P^x$. Thus $gx^{-1} \in N_G(P)$ and so $g \in N_G(P)x$. Since g was an arbitrary element of G, this proves $G = N_G(P)H$.

Apply the Second Isomorphism Theorem to $G = N_G(P)H$ to conclude that

$$|G : H| = |N_G(P) : N_G(P) \cap H|$$

so $|G : H|$ divides $|N_G(P)|$, completing the proof.

Proposition 7. A finite group is nilpotent if and only if every maximal subgroup is normal.

Proof: Let G be a finite nilpotent group and let M be a maximal subgroup of G. As in the proof of Theorem 1, since $M < N_G(M)$ (by Theorem 3(2)) maximality of M forces $N_G(M) = G$, i.e. $M \trianglelefteq G$.

Conversely, assume every maximal subgroup of the finite group G is normal. Let P be a Sylow p-subgroup of G. We prove $P \trianglelefteq G$ and conclude that G is nilpotent by Theorem 3(3). If P is not normal in G let M be a maximal subgroup of G containing $N_G(P)$. By hypothesis, $M \trianglelefteq G$ hence by Frattini's Argument $G = MN_G(P)$. Since $N_G(P) \leq M$ we have $MN_G(P) = M$, a contradiction. This establishes the converse.

Commutators and the Lower Central Series

For the sake of completeness we include the definition of the *lower central series* of a group and state its relation to the upper central series. Since we shall not be using these results in future, the proofs are left as (straightforward) exercises.

Recall that the commutator of two elements x, y in a group G is defined as

$$[x, y] = x^{-1}y^{-1}xy,$$

and the commutator of two subgroups H and K of G is

$$[H, K] = \langle [h, k] \mid h \in H,\ k \in K \rangle.$$

Basic properties of commutators and the commutator subgroup were established in Section 5.4.

Definition. For any (finite or infinite) group G define the following subgroups inductively:

$$G^0 = G, \qquad G^1 = [G, G] \quad \text{and} \quad G^{i+1} = [G, G^i].$$

The chain of groups

$$G^0 \geq G^1 \geq G^2 \geq \cdots$$

is called the *lower central series of* G. (The term "lower" indicates that $G^i \geq G^{i+1}$).

As with the upper central series we include in the exercises at the end of this section the verification that G^i is a characteristic subgroup of G for all i. The next theorem shows the relation between the upper and lower central series of a group.

Theorem 8. A group G is nilpotent if and only if $G^n = 1$ for some $n \geq 0$. More precisely, G is nilpotent of class c if and only if c is the smallest nonnegative integer such that $G^c = 1$. If G is nilpotent of class c then

$$Z_i(G) \leq G^{c-i-1} \leq Z_{i+1}(G) \quad \text{for all } i \in \{0, 1, \ldots, c-1\}.$$

Proof: This is proved by a straightforward induction on the length of either the upper or lower central series.

The terms of the upper and lower central series do not necessarily coincide in general although in some groups this does occur.

Remarks

(1) If G is abelian, we have already seen that $G' = G^1 = 1$ so the lower central series terminates in the identity after one term.

(2) As with the upper central series, for any finite group there must, by order considerations, be an integer n such that

$$G^n = G^{n+1} = G^{n+2} = \cdots.$$

For non-nilpotent groups, G^n is a nontrivial subgroup of G. For example, in Section 5.4 we showed that $S_3' = S_3^1 = A_3$. Since S_3 is not nilpotent, we must have $S_3^2 = A_3$. In fact

$$(123) = [(12), (132)] \in [S_3, S_3^1] = S_3^2.$$

Once two terms in the lower central series are the same, the chain stabilizes at that point i.e. all terms thereafter are equal to these two. Thus $S_3^i = A_3$ for all $i \geq 2$. Note that S_3 is an example where the lower central series has two distinct terms whereas all terms in the upper central series are equal to the identity (in particular, for non-nilpotent groups these series need not have the same length).

Solvable Groups and the Derived Series

Recall that in Section 3.4 a solvable group was defined as one possessing a series:

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_s = G$$

such that each factor H_{i+1}/H_i is abelian. We now give another characterization of solvability in terms of a descending series of characteristic subgroups.

Definition. For any group G define the following sequence of subgroups inductively:

$$G^{(0)} = G, \qquad G^{(1)} = [G, G] \quad \text{and} \quad G^{(i+1)} = [G^{(i)}, G^{(i)}] \quad \text{for all } i \geq 1.$$

This series of subgroups is called the *derived* or *commutator* series of G.

The terms of this series are also often written as: $G^{(1)} = G'$, $G^{(2)} = G''$, etc. Again it is left as an exercise to show that each $G^{(i)}$ is characteristic in G for all i.

It is important to note that although $G^{(0)} = G^0$ and $G^{(1)} = G^1$, it is not in general true that $G^{(i)} = G^i$. The difference is that the definition of the $i+1^{\text{st}}$ term in the lower central series is the commutator of the ith term with the *whole* group G whereas the $i+1^{\text{st}}$ term in the derived series is the commutator of the ith term with itself. Hence

$$G^{(i)} \leq G^i \quad \text{for all } i$$

and the containment can be proper. For example, in $G = S_3$ we have already seen that $G^1 = G' = A_3$ and $G^2 = [S_3, A_3] = A_3$, whereas $G^{(2)} = [A_3, A_3] = 1$ (A_3 being abelian).

Theorem 9. A group G is solvable if and only if $G^{(n)} = 1$ for some $n \geq 0$.

Proof: Assume first that G is solvable and so possesses a series

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_s = G$$

such that each factor H_{i+1}/H_i is abelian. We prove by induction that $G^{(i)} \leq H_{s-i}$. This is true for $i = 0$, so assume $G^{(i)} \leq H_{s-i}$. Then

$$G^{(i+1)} = [G^{(i)}, G^{(i)}] \leq [H_{s-i}, H_{s-i}].$$

Since H_{s-i}/H_{s-i-1} is abelian, by Proposition 5.7(4), $[H_{s-i}, H_{s-i}] \leq H_{s-i-1}$. Thus $G^{(i+1)} \leq H_{s-i-1}$, which completes the induction. Since $H_0 = 1$ we have $G^{(s)} = 1$.

Conversely, if $G^{(n)} = 1$ for some $n \geq 0$, Proposition 5.7(4) shows that if we take H_i to be $G^{(n-i)}$ then H_i is a normal subgroup of H_{i+1} with abelian quotient, so the derived series itself satisfies the defining condition for solvability of G. This completes the proof.

If G is solvable, the smallest nonnegative n for which $G^{(n)} = 1$ is called the *solvable length* of G. The derived series is a series of shortest length whose successive quotients are abelian and it has the additional property that it consists of subgroups that are characteristic in the *whole* group (as opposed to each just being normal in the *next* in the initial definition of solvability). Its "intrinsic" definition also makes it easier to work with in many instances, as the following proposition (which reproves some results and exercises from Section 3.4) illustrates:

Proposition 10. Let G and K be groups, let H be a subgroup of G and let $\varphi : G \to K$ be a surjective homomorphism.

(1) $H^{(i)} \leq G^{(i)}$ for all $i \geq 0$. In particular, if G is solvable, then so is H, i.e. subgroups of solvable groups are solvable (and the solvable length of H is less than or equal to the solvable length of G).

(2) $\varphi(G^{(i)}) = K^{(i)}$. In particular, homomorphic images and quotient groups of solvable groups are solvable (of solvable length less than or equal to that of the domain group).

(3) If N is normal in G and both N and G/N are solvable then so is G.

Proof: Part (1) follows from the observation that since $H \leq G$, by definition of commutator subgroups, $[H, H] \leq [G, G]$, i.e. $H^{(1)} \leq G^{(1)}$. Then, by induction,

$$H^{(i)} \leq G^{(i)} \quad \text{for all } i \in \mathbb{Z}^+.$$

In particular, if $G^{(n)} = 1$ for some n, then also $H^{(n)} = 1$. This establishes (1).

To prove (2) note that by definition of commutators,

$$\varphi([x, y]) = [\varphi(x), \varphi(y)]$$

so by induction $\varphi(G^{(i)}) \leq K^{(i)}$. Since φ is surjective, every commutator in K is the image of a commutator in G, hence again by induction we obtain equality for all i. Again, if $G^{(n)} = 1$ for some n then $K^{(n)} = 1$. This proves (2).

Finally, if G/N and N are solvable, of lengths n and m respectively then by (2) applied to the natural projection $\varphi : G \to G/N$ we obtain

$$\varphi(G^{(n)}) = (G/N)^{(n)} = 1N$$

i.e. $G^{(n)} \leq N$. Thus $G^{(n+m)} = (G^{(n)})^{(m)} \leq N^{(m)} = 1$. Theorem 9 shows that G is solvable, which completes the proof.

Some additional conditions under which finite groups are solvable are the following:

Theorem 11. Let G be a finite group.

(1) (Burnside) If $|G| = p^a q^b$ for some primes p and q, then G is solvable.

(2) (Philip Hall) If for every prime p dividing $|G|$ we factor the order of G as $|G| = p^a m$ where $(p, m) = 1$, and G has a subgroup of order m, then G is solvable (i.e. if for all primes p, G has a subgroup whose index equals the order of a Sylow p-subgroup, then G is solvable — such subgroups are called Sylow p-complements).

(3) (Feit–Thompson) If $|G|$ is odd then G is solvable.

(4) (Thompson) If for every pair of elements $x, y \in G$, $\langle x, y \rangle$ is a solvable group, then G is solvable.

We shall prove Burnside's Theorem in Chapter 16 and deduce Philip Hall's generalization of it. As mentioned in Section 3.5, the proof of the Feit–Thompson Theorem takes 255 pages. Thompson's Theorem is a consequence of a 475 page paper (that in turn relies ultimately on the Feit–Thompson Theorem).

A Proof of the Fundamental Theorem of Finite Abelian Groups

We sketch a group-theoretic proof of the result that every finite abelian group is a direct product of cyclic groups (i.e. Parts 1 and 2 of Theorem 5, Section 5.2) – the Classification of Finitely Generated Abelian Groups (Theorem 3, Section 5.2) will be derived as a consequence of a more general theorem in Chapter 12.

By Corollary 4 it suffices to prove that for p a prime, any abelian p-group is a direct product of cyclic groups (the divisibility condition in Theorem 5.5 is trivially achieved by reordering factors). Let A be an abelian p-group. We proceed by induction on $|A|$.

If E is an elementary abelian p-group (i.e. $x^p = 1$ for all $x \in E$), we first prove the following result:

$$\text{for any } x \in E, \text{ there exists } M \leq E \text{ with } E = M \times \langle x \rangle.$$

If $x = 1$, let $M = E$. Otherwise let M be a subgroup of E of maximal order subject to the condition that x not be an element of M. If M is not of index p in E, let $\overline{E} = E/M$. Then $\overline{E}$ is elementary abelian and there exists $\overline{y} \in \overline{E} - \langle \overline{x} \rangle$. Since $\overline{y}$ has order p, we also have $\overline{x} \notin \langle \overline{y} \rangle$. The complete preimage of $\langle \overline{y} \rangle$ in E is a subgroup of E that does not contain x and whose order is larger than the order of M, contrary to the choice of M. This proves $|E : M| = p$, hence

$$E = M\langle x \rangle \quad \text{and} \quad M \cap \langle x \rangle = 1.$$

By the recognition theorem for direct products, Theorem 5.9, $E = M \times \langle x \rangle$, as asserted.

Now let $\varphi : A \to A$ be defined by $\varphi(x) = x^p$ (see Exercise 7, Section 5.2). Then φ is a homomorphism since A is abelian. Denote the kernel of φ by K and denote the image of φ by H. By definition $K = \{x \in A \mid x^p = 1\}$ and H is the subgroup of A consisting of pth powers. Note that both K and A/H are elementary abelian. By the First Isomorphism Theorem

$$|A : H| = |K|.$$

By induction,

$$\begin{aligned} H &= \langle h_1 \rangle \times \cdots \times \langle h_r \rangle \\ &\cong Z_{p^{\alpha_1}} \times \cdots \times Z_{p^{\alpha_r}} \quad \alpha_i \geq 1,\ i = 1, 2, \ldots, r. \end{aligned}$$

By definition of φ, there exist elements $g_i \in A$ such that $g_i^p = h_i$, $1 \leq i \leq r$. Let $A_0 = \langle g_1, \cdots, g_r \rangle$. It is an exercise to see that

(a) $A_0 = \langle g_1 \rangle \times \cdots \times \langle g_r \rangle$,

(b) $A_0/H = \langle g_1 H \rangle \times \cdots \times \langle g_r H \rangle$ is elementary abelian of order p^r, and

(c) $H \cap K = \langle h_1^{p^{\alpha_1 - 1}} \rangle \times \cdots \times \langle h_r^{p^{\alpha_r - 1}} \rangle$ is elementary abelian of order p^r.

If K is contained in H, then $|K| = |K \cap H| = p^r = |A_0 : H|$. In this case by comparing orders we see that $A_0 = A$ and the theorem is proved. Assume therefore that K is not a subgroup of H and use the bar notation to denote passage to the

quotient group A/H. Let $x \in K - H$, so $|\overline{x}| = |x| = p$. By the initial remark of the proof applied to the elementary abelian p-group $E = \overline{A}$, there is a subgroup $\overline{M}$ of $\overline{A}$ such that

$$\overline{A} = \overline{M} \times \langle \overline{x} \rangle.$$

If M is the complete preimage in A of $\overline{M}$, then since x has order p and $x \notin M$, we have $\langle x \rangle \cap M = 1$. By the recognition theorem for direct products,

$$A = M \times \langle x \rangle.$$

By induction, M is a direct product of cyclic groups, hence so is A. This completes the proof.

The uniqueness of the decomposition of a finite abelian group into a direct product of cyclic groups (Part 3 of Theorem 5.5) can also be proved by induction using the pth-power map (i.e. using Exercise 7, Section 5.2). This is essentially the procedure we follow in Section 12.1 for the uniqueness part of the proof of the Fundamental Theorem of Finitely Generated Abelian Groups.

EXERCISES

1. Prove that $Z_i(G)$ is a characteristic subgroup of G for all i.
2. Prove Parts 2 and 4 of Theorem 1 for G a finite nilpotent group, not necessarily a p-group.
3. Prove that a finite nilpotent group of order n has a normal subgroup of order m for each divisor m of n.
4. Prove that a maximal subgroup of a finite nilpotent group has prime index.
5. Prove Parts 2 and 4 of Theorem 1 for G an infinite nilpotent group.
6. Show that if $G/Z(G)$ is nilpotent then G is nilpotent.
7. Prove that subgroups and quotient groups of nilpotent groups are nilpotent (your proof should work for infinite groups). Give an explicit example of a group G which possesses a normal subgroup H such that both H and G/H are nilpotent but G is not nilpotent.
8. Prove that if p is a prime and P is a non-abelian group of order p^3 then $|Z(P)| = p$ and $P/Z(P) \cong Z_p \times Z_p$.
9. Prove that a finite group G is nilpotent if and only if whenever $a,\ b \in G$ with $(|a|, |b|) = 1$ then $ab = ba$. [Use Part 4 of Theorem 3.]
10. Prove that D_{2n} is nilpotent if and only if n is a power of 2. [Use Exercise 9.]
11. Give another proof of Proposition 5 under the additional assumption that G is abelian by invoking the Fundamental Theorem of Finite Abelian Groups.
12. Find the upper and lower central series for A_4 and S_4.
13. Find the upper and lower central series for A_n and S_n, $n \geq 5$.
14. Prove that G^i is a characteristic subgroup of G for all i.
15. Prove that $Z_i(D_{2^n}) = D_{2^n}^{n-1-i}$.
16. Prove that $\mathbb{Q}$ has no maximal subgroups. [Recall Exercise 21, Section 3.2.]
17. Prove that $G^{(i)}$ is a characteristic subgroup of G for all i.
18. Show that if G'/G'' and G''/G''' are both cyclic then $G'' = 1$. [You may assume $G''' = 1$. Then G/G'' acts by conjugation on the cyclic group G''.]

19. Show that there is no group whose commutator subgroup is isomorphic to S_4. [Use the preceding exercise.]

20. Let p be a prime, let P be a p-subgroup of the finite group G, let N be a normal subgroup of G whose order is relatively prime to p and let $\overline{G} = G/N$. Prove the following:
    (a) $N_{\overline{G}}(\overline{P}) = \overline{N_G(P)}$ [Use Frattini's Argument.]
    (b) $C_{\overline{G}}(\overline{P}) = \overline{C_G(P)}$. [Use part (a).]

For any group G the *Frattini subgroup* of G (denoted by $\Phi(G)$) is defined to be the intersection of all the maximal subgroups of G (if G has no maximal subgroups, set $\Phi(G) = 1$). The next few exercises deal with this important subgroup.

21. Prove that $\Phi(G)$ is a characteristic subgroup of G.

22. Prove that if $N \trianglelefteq G$ then $\Phi(N) \leq \Phi(G)$. Give an explicit example where this containment does not hold if N is not normal in G.

23. Compute $\Phi(S_3)$, $\Phi(A_4)$, $\Phi(S_4)$, $\Phi(A_5)$ and $\Phi(S_5)$.

24. Say an element x of G is a *nongenerator* if for every proper subgroup H of G, $\langle x, H\rangle$ is also a proper subgroup of G. Prove that $\Phi(G)$ is the set of nongenerators of G (here $|G| > 1$).

25. Let G be a finite group. Prove that $\Phi(G)$ is nilpotent. [Use Frattini's Argument to prove that every Sylow subgroup of $\Phi(G)$ is normal in G.]

26. Let p be a prime, let P be a finite p-group and let $\overline{P} = P/\Phi(P)$.
    (a) Prove that $\overline{P}$ is an elementary abelian p-group. [Show that $P' \leq \Phi(P)$ and that $x^p \in \Phi(P)$ for all $x \in P$.]
    (b) Prove that if N is any normal subgroup of P such that P/N is elementary abelian then $\Phi(P) \leq N$. State this (universal) property in terms of homomorphisms and commutative diagrams.
    (c) Let $\overline{P}$ be elementary abelian of order p^r (by (a)). Deduce from Exercise 24 that if $\overline{x_1}, \overline{x_2}, \ldots, \overline{x_r}$ are any basis for the r-dimensional vector space $\overline{P}$ over $\mathbb{F}_p$ and if x_i is any element of the coset $\overline{x_i}$, then $P = \langle x_1, x_2, \ldots, x_r\rangle$. Show conversely that if $y_1, y_2, \ldots, y_s$ is any set of generators for P, then $s \geq r$ (you may assume that every minimal generating set for an r-dimensional vector space has r elements, i.e. every basis has r elements). Deduce *Burnside's Basis Theorem*: a set $y_1, \ldots, y_s$ is a minimal generating set for P if and only if $\overline{y_1}, \ldots, \overline{y_s}$ is a basis of $\overline{P} = P/\Phi(P)$. Deduce that any minimal generating set for P has r elements.
    (d) Prove that if $P/\Phi(P)$ is cyclic then P is cyclic. Deduce that if P/P' is cyclic then so is P.
    (e) Let σ be any automorphism of P of prime order q with $q \neq p$. Show that if σ fixes the coset $x\Phi(P)$ then σ fixes some element of this coset (note that since $\Phi(P)$ is characteristic in P every automorphism of P induces an automorphism of $P/\Phi(P)$). [Use the observation that σ acts a permutation of order 1 or q on the p^a elements in the coset $x\Phi(P)$.]
    (f) Use parts (e) and (c) to deduce that every nontrivial automorphism of P of order prime to p induces a nontrivial automorphism on $P/\Phi(P)$. Deduce that any group of automorphisms of P which has order prime to p is isomorphic to a subgroup of $\mathrm{Aut}(\overline{P}) = GL_r(\mathbb{F}_p)$.

27. Generalize part (d) of the preceding exercise as follows: let p be a prime, let P be a p-group and let $\overline{P} = P/\Phi(P)$ be elementary abelian of order p^r. Prove that P has exactly $\dfrac{p^r - 1}{p - 1}$

maximal subgroups. [Since every maximal subgroup of P contains $\Phi(P)$, the maximal subgroups of P are, by the Lattice Isomorphism Theorem, in bijective correspondence with the maximal subgroups of the elementary abelian group $\overline{P}$. It therefore suffices to show that the number of maximal subgroups of an elementary abelian p-group of order p^r is as stated above. One way of doing this is to use the result that an abelian group is isomorphic to its dual group (cf. Exercise 14 in Section 5.2) so the number of subgroups of *index* p equals the number of subgroups of *order* p.]

28. Prove that if p is a prime and $P = Z_p \times Z_{p^2}$ then $|\Phi(P)| = p$ and $P/\Phi(P) \cong Z_p \times Z_p$. Deduce that P has $p+1$ maximal subgroups.

29. Prove that if p is a prime and P is a non-abelian group of order p^3 then $\Phi(P) = Z(P)$ and $P/\Phi(P) \cong Z_p \times Z_p$. Deduce that P has $p+1$ maximal subgroups.

30. Let p be an odd prime, let $P_1 = Z_p \times Z_{p^2}$ and let P_2 be the non-abelian group of order p^3 which has an element of order p^2. Prove that P_1 and P_2 have the same lattice of subgroups.

31. For any group G a *minimal normal subgroup* is a normal subgroup M of G such that the only normal subgroups of G which are contained in M are 1 and M. Prove that every minimal normal subgroup of a finite solvable group is an elementary abelian p-group for some prime p. [If M is a minimal normal subgroup of G, consider its characteristic subgroups: M' and $\langle x^p \mid x \in M\rangle$.]

32. Prove that every maximal subgroup of a finite solvable group has prime power index. [Let H be a maximal subgroup of G and let M be a minimal normal subgroup of G – cf. the preceding exercise. Apply induction to G/M and consider separately the two cases: $M \leq H$ and $M \not\leq H$.]

33. Let π be any set of primes. A subgroup H of a finite group is called a *Hall π-subgroup* of G if the only primes dividing $|H|$ are in the set π and $|H|$ is relatively prime to $|G : H|$. (Note that if $\pi = \{p\}$, Hall π-subgroups are the same as Sylow p-subgroups. Hall subgroups were introduced in Exercise 10 of Section 3.3). Prove the following generalization of Sylow's Theorem for solvable groups: if G is a finite solvable group then for every set π of primes, G has a Hall π- subgroup and any two Hall π- subgroups (for the same set π) are conjugate in G. [Fix π and proceed by induction on $|G|$, proving both existence and conjugacy at once. Let M be a minimal normal subgroup of G, so M is a p-group for some prime p. If $p \in \pi$, apply induction to G/M. If $p \notin \pi$, reduce to the case $|G| = p^\alpha n$, where $p^\alpha = |M|$ and n is the order of a Hall π- subgroup of G. In this case let N/M be a minimal normal subgroup of G/M, so N/M is a q-group for some prime $q \neq p$. Let $Q \in Syl_q(N)$. If $Q \trianglelefteq G$ argue as before with Q in place of M. If Q is not normal in G, use Frattini's Argument to show $N_G(Q)$ is a Hall π- subgroup of G and establish conjugacy in this case too.]

The following result shows how to produce normal p-subgroups of some groups on which the elements of order prime to p act faithfully by conjugation. Exercise 26(f) then applies to restrict these actions and give some information about the structure of the group.

34. Let p be a prime dividing the order of the finite solvable group G. Assume G has no non-trivial normal subgroups of order prime to p. Let P be the largest normal p-subgroup of G (cf. Exercise 37, Section 4.5). Note that Exercise 31 above shows that $P \neq 1$. Prove that $C_G(P) \leq P$, i.e. $C_G(P) = Z(P)$. [Let $N = C_G(P)$ and use the preceding exercise to show $N = Z(P) \times H$ for some Hall π- subgroup H of N – here π is the set of all prime divisors of $|N|$ except for p. Show $H \trianglelefteq G$ to obtain the desired conclusion: $H = 1$.]

35. Prove that if G is a finite group in which every proper subgroup is nilpotent, then G is solvable. [Show that a minimal counterexample is simple. Let M and N be distinct maximal subgroups chosen with $|M \cap N|$ as large as possible and apply Part 2 of Theorem 3 to show that $M \cap N = 1$. Now apply the methods of Exercise 53 in Section 4.5.]

36. Let p be a prime, let V be a nonzero finite dimensional vector space over the field of p elements and let φ be an element of $GL(V)$ of order a power of p (i.e. V is a nontrivial elementary abelian p-group and φ is an automorphism of V of p-power order). Prove that there is some nonzero element $v \in V$ such that $\varphi(v) = v$, i.e. φ has a nonzero fixed point on V.

37. Let V be a finite dimensional vector space over the field of 2 elements and let φ be an element of $GL(V)$ of order 2. (i.e. V is a nontrivial elementary abelian 2-group and φ is an automorphism of V of order 2). Prove that the map $v \mapsto v + \varphi(v)$ is a homomorphism from V to itself. Show that every element in the image of this map is fixed by φ. Deduce that the subspace of elements of V which are fixed by φ has dimension $\geq \frac{1}{2}(\text{dimension } V)$. (Note that if G is the semidirect product of V with $\langle\varphi\rangle$, where $V \trianglelefteq G$ and φ acts by conjugation on V by sending each $v \in V$ to $\varphi(v)$, then the fixed points of φ on V are $C_V(\varphi)$ and the above map is simply the commutator map: $v \mapsto [v, \varphi]$. In this terminology the problem is to show that $|C_V(\varphi)|^2 \geq |V|$.)

38. Use the preceding exercise to prove that if P is a 2-group which has a cyclic center and M is a subgroup of index 2 in P, then the center of M has rank ≤ 2. [The group G/M of order 2 acts by conjugation on the $\mathbb{F}_2$ vector space: $\{z \in Z(M) \mid z^2 = 1\}$ and the fixed points of this action are in the center of P.]

6.2 APPLICATIONS IN GROUPS OF MEDIUM ORDER

The purpose of this section is to work through a number of examples which illustrate many of the techniques we have developed. These examples use Sylow's Theorems extensively and demonstrate how they are applied in the study of finite groups. Motivated by the Hölder Program we address primarily the problem of showing that for certain n every group of order n has a proper, nontrivial normal subgroup (i.e. there are no simple groups of order n). In most cases we shall stop once this has been accomplished. However readers should be aware that in the process of achieving this result we shall already have determined a great deal of information about arbitrary groups of given order n for the n that we consider. This information could be built upon to classify groups of these orders (but in general this requires techniques beyond the simple use of semidirect products to construct groups).

Since for p a prime we have already proved that there are no simple p-groups (other than the cyclic group of order p, Z_p) and since the structure of p-groups can be very complicated (recall the table in Section 5.3), we shall not study the structure of p-groups explicitly. Rather, the theory of p-groups developed in the preceding section will be applied to subgroups of groups of non-prime-power order.

Finally, for certain n (e.g. 60, 168, 360, 504, ...) there do exist simple groups of order n so, of course, we cannot force every group of these orders to be non-simple. As in Section 4.5 we can, in certain cases, prove there is a unique simple group of order n and unravel some of its internal structure (Sylow numbers, etc.). We shall

study simple groups of order 168 as an additional test case. Thus the Sylow Theorems will be applied in a number of different contexts to show how groups of a given order may be manipulated.

We shall end this section with some comments on the existence problem for groups, particularly for finite simple groups.

For $n < 10,000$ there are 60 odd, non-prime-power numbers for which the congruence conditions of Sylow's Theorems do *not* force at least one of the Sylow subgroups to be normal i.e. n_p can be > 1 for all primes $p \mid n$ (recall that n_p denotes the number of Sylow p-subgroups). For example, no numbers of the form pq, where p and q are distinct primes occur in our list by results of Section 4.5. In contrast, for even numbers < 500 there are already 46 candidates for orders of simple groups (the congruence conditions allow many more possibilities). Many of our numerical examples arise from these lists of numbers and we often use odd numbers because the Sylow congruence conditions allow fewer values for n_p. The purpose of these examples is to illustrate the use of the results we have proved. Many of these examples can be dealt with by more advanced techniques (for example, the Feit–Thompson Theorem proves that there are *no* simple groups of odd composite order).

As we saw in the case $n = 30$ in Section 4.5, even though Sylow's Theorem permitted $n_5 = 6$ and $n_3 = 10$, further examination showed that any group of order 30 must have both $n_5 = 1$ and $n_3 = 1$. Thus the congruence part of Sylow's Theorem is a sufficient but by no means necessary condition for normality of a Sylow subgroup. For many n (e.g. $n = 120$) we can prove that there are no simple groups of order n, so there is a nontrivial normal subgroup but this subgroup may not be a Sylow subgroup. For example, S_5 and $SL_2(\mathbb{F}_5)$ both have order 120. The group S_5 has a unique nontrivial proper normal subgroup of order 60 (A_5) and $SL_2(\mathbb{F}_5)$ has a unique nontrivial proper normal subgroup of order 2 ($Z(SL_2(\mathbb{F}_5)) \cong Z_2$), neither of which is a Sylow subgroup. Our techniques for producing normal subgroups must be flexible enough to cover such diverse possibilities. In this section we shall examine Sylow subgroups for different primes dividing n, intersections of Sylow subgroups, normalizers of p-subgroups and many other less obvious subgroups. The elementary methods we outline are by no means exhaustive, even for groups of "medium" order.

Some Techniques

Before listing some techniques for producing normal subgroups in groups of a given ("medium") order we note that in all the problems where one deals with groups of order n, for some specific n, it is first necessary to factor n into prime powers and then to compute the permissible values of n_p, for all primes p dividing n. We emphasize the need to be comfortable computing mod p when carrying out the last step. The techniques we describe may be listed as:

(1) Counting elements.
(2) Exploiting subgroups of small index.
(3) Permutation representations.
(4) Playing p-subgroups off against each other for different primes p.
(5) Studying normalizers of intersections of Sylow p-subgroups.

Counting Elements

Let G be a group of order n, let p be a prime dividing n and let $P \in Syl_p(G)$. If $|P| = p$, then every nonidentity element of P has order p and every element of G of order p lies in some conjugate of P. By Lagrange's Theorem distinct conjugates of P intersect in the identity, hence in this case the number of elements of G of order p is $n_p(p-1)$.

If Sylow p-subgroups for different primes p have prime order and we assume none of these is normal, we can sometimes show that the number of elements of prime order is $> |G|$. This contradiction would show that at least one of the n_p's must be 1 (i.e. some Sylow subgroup is normal in G).

This is the argument we used (in Section 4.5) to prove that there are no simple groups of order 30. For another example, suppose $|G| = 105 = 3 \cdot 5 \cdot 7$. If G were simple, we must have $n_3 = 7$, $n_5 = 21$ and $n_7 = 15$. Thus

$$\begin{array}{ll} \text{the number of elements of order 3 is } 7 \cdot 2 & = 14 \\ \text{the number of elements of order 5 is } 21 \cdot 4 & = 84 \\ \text{the number of elements of order 7 is } 15 \cdot 6 & = 90 \\ \hline \text{the number of elements of prime order is 188} & > |G|. \end{array}$$

Sometimes counting elements of prime order does not lead to too many elements. However, there may be so few elements remaining that there must be a normal subgroup involving these elements. This was (in essence) the technique used in Section 4.5 to show that in a group of order 12 either $n_2 = 1$ or $n_3 = 1$. This technique works particularly well when G has a Sylow p-subgroup P of order p such that $N_G(P) = P$. For example, let $|G| = 56$. If G were simple, the only possibility for the number of Sylow 7-subgroups is 8, so

$$\text{the number of elements of order 7 is } 8 \cdot 6 = 48.$$

Thus there are $56 - 48 = 8$ elements remaining in G. Since a Sylow 2-subgroup contains 8 elements (none of which have order 7), there can be at most one Sylow 2-subgroup, hence G has a normal Sylow 2-subgroup.

Exploiting Subgroups of Small Index

Recall that the results of Section 4.2 show that if G has a subgroup H of index k, then there is a homomorphism from G into the symmetric group S_k whose kernel is contained in H. If $k > 1$, this kernel is a proper normal subgroup of G and if we are trying to prove that G is not simple, we may, by way of contradiction, assume that this kernel is the identity. Then, by the First Isomorphism Theorem, G is isomorphic to a subgroup of S_k. In particular, the order of G divides $k!$. This argument shows that if k is the smallest integer with $|G|$ dividing $k!$ for a finite simple group G then G contains no proper subgroups of index less than k. This smallest permissible index k should be calculated at the outset of the study of groups of a given order n. In the examples we

consider this is usually quite easy: n will often factor as

$$p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_s^{\alpha_s} \quad \text{with} \quad p_1 < p_2 < \cdots < p_s$$

and α_s is usually equal to 1 or 2 in our examples. In this case the minimal index of a proper subgroup will have to be at least p_s (respectively $2p_s$) and this is often its exact value.

For example, there is no simple group of order 3393, because if $n = 3393 = 3^2 \cdot 13 \cdot 29$, then the minimal index of a proper subgroup is 29 (n does not divide 28! because 29 does not divide 28!). However any simple group of order 3393 must have $n_3 = 13$, so for $P \in Syl_3(G)$, $N_G(P)$ has index 13, a contradiction.

Permutation Representations

This method is a refinement of the preceding one. As above, if G is a simple group of order n with a proper subgroup of index k, then G is isomorphic to a subgroup of S_k. We may identify G with this subgroup and so assume $G \leq S_k$. Rather than relying only on Lagrange's Theorem for our contradiction (this was what we did for the preceding technique) we can sometimes show by calculating within S_k that S_k contains no simple subgroup of order n. Some restrictions which may enable one to show such a result are:

(1) if G contains an element or subgroup of a particular order, so must S_k, and
(2) if $P \in Syl_p(G)$ and if P is also a Sylow p-subgroup of S_k, then $|N_G(P)|$ must divide $|N_{S_k}(P)|$.

Condition (2) arises frequently when p is a prime, $k = p$ or $p + 1$ and G has a subgroup of index k. In this case p^2 does not divide k!, so Sylow p-subgroups of G are also Sylow p-subgroups of S_k. Since now Sylow p-subgroups of S_k are precisely the groups generated by a p-cycle, and distinct Sylow p-subgroups intersect in the identity,

$$\text{the number of Sylow } p\text{-subgroups of } S_k = \frac{\text{the number of } p\text{-cycles}}{\text{the number of } p\text{-cycles in a Sylow } p\text{-subgro}}$$

$$= \frac{k \cdot (k-1) \cdots (k-p+1)}{p(p-1)}.$$

This number gives the index in S_k of the normalizer of a Sylow p-subgroup of S_k. Thus for $k = p$ or $p + 1$

$$|N_{S_k}(P)| = p(p-1) \qquad (k = p \text{ or } k = p+1)$$

(cf. also the corresponding discussion for centralizers of elements in symmetric groups in Section 4.3 and the last exercises in Section 4.3). This proves, under the above hypotheses, that $|N_G(P)|$ must divide $p(p-1)$.

For example, if G were a simple group of order $396 = 2^2 \cdot 3^2 \cdot 11$, we must have $n_{11} = 12$, so if $P \in Syl_{11}(G)$, $|G : N_G(P)| = 12$ and $|N_G(P)| = 33$. Since G has a subgroup of index 12, G is isomorphic to a subgroup of S_{12}. But then (considering G as actually contained in S_{12}) $P \in Syl_{11}(S_{12})$ and $|N_{S_{12}}(P)| = 110$. Since $N_G(P) \leq$

$N_{S_{12}}(P)$, this would imply $33 \mid 110$, clearly impossible, so we cannot have a simple group of order 396.

We can sometimes squeeze a little bit more out of this method by working in A_k rather than S_k. This slight improvement helps only occasionally and only for groups of even order. It is based on the following observations (the first of which we have made earlier in the text)

Proposition 12

(1) If G has no subgroup of index 2 and $G \leq S_k$, then $G \leq A_k$.

(2) If $P \in Syl_p(S_k)$ for some odd prime p, then $P \in Syl_p(A_k)$ and $|N_{A_k}(P)| = \frac{1}{2}|N_{S_k}(P)|$.

Proof: The first assertion follows from the Second Isomorphism Theorem: if G is not contained in A_k, then $A_k < GA_k$ so we must have $GA_k = S_k$. But now

$$2 = |S_k : A_k| = |GA_k : A_k| = |G : G \cap A_k|$$

so G has a subgroup, $G \cap A_k$, of index 2.

To prove (2) note that if $P \in Syl_p(S_k)$, for some odd prime p, by (1) (or order considerations) $P \leq A_k$, hence $P \in Syl_p(A_k)$ as well. By Frattini's Argument (Proposition 6)

$$S_k = N_{S_k}(P)A_k$$

so, in particular, $N_{S_k}(P)$ is not contained in A_k. This forces $N_{S_k}(P) \cap A_k$ $(= N_{A_k}(P))$ to be a subgroup of index 2 in $N_{S_k}(P)$.

For example, there is no simple group of order 264. Suppose G were a simple group of order $264 = 2^3 \cdot 3 \cdot 11$. We must have $n_{11} = 12$. As usual, G would be isomorphic to a subgroup of S_{12}. Since G is simple (hence contains no subgroup of index 2), $G \leq A_{12}$. Let $P \in Syl_{11}(G)$. Since $n_{11} = 12 = |G : N_G(P)|$, we have $|N_G(P)| = 22$. As above,

$$|N_{A_{12}}(P)| = \tfrac{1}{2}|N_{S_{12}}(P)| = \tfrac{1}{2}11(11-1) = 55;$$

however, 22 does not divide 55, a contradiction to $N_G(P) \leq N_{A_{12}}(P)$.

Finally, we emphasize that we have only barely touched upon the combinatorial information available from certain permutation representations. Whenever possible in the remaining examples we shall illustrate other applications of this technique.

Playing p-Subgroups Off Against Each Other for Different Primes p

Suppose p and q are distinct primes such that every group of order pq is cyclic. This is equivalent to $p \nmid q-1$, where $p < q$. If G has a Sylow q-subgroup Q of order q and $p \mid |N_G(Q)|$, applying Cauchy's Theorem in $N_G(Q)$ gives a group P of order p normalizing Q (note that P need not be a Sylow p-subgroup of G). Thus PQ is a group and if PQ is abelian, we obtain

$$PQ \leq N_G(P) \quad \text{and so} \quad q \mid |N_G(P)|.$$

(A symmetric argument applies if Sylow p-subgroups of G have order p and q divides the order of a Sylow p-normalizer). This numerical information alone may be sufficient to force $N_G(P) = G$ (i.e. $P \trianglelefteq G$), or at least to force $N_G(P)$ to have index smaller than the minimal index permitted by permutation representations, giving a contradiction by a preceding technique.

For example, there are no simple groups of order 1785. If there were, let G be a simple group of order $1785 = 3 \cdot 5 \cdot 7 \cdot 17$. The only possible value for n_{17} is 35, so if Q is a Sylow 17-subgroup, $|G : N_G(Q)| = 35$. Thus $|N_G(Q)| = 3 \cdot 17$. Let P be a Sylow 3-subgroup of $N_G(Q)$. The group PQ is abelian since 3 does not divide $17 - 1$, so $Q \leq N_G(P)$ and $17 \mid |N_G(P)|$. In this case $P \in Syl_3(G)$. The permissible values of n_3 are 7, 85 and 595; however, since $17 \mid |N_G(P)|$, we cannot have $17 \mid |G : N_G(P)| = n_3$. Thus $n_3 = 7$. But G has no proper subgroup of index < 17 (the minimal index of a proper subgroup is 17 for this order), a contradiction. Alternatively, if $n_3 = 7$, then $|N_G(P)| = 3 \cdot 5 \cdot 17$, and by Sylow's Theorem applied in $N_G(P)$ we have $Q \trianglelefteq N_G(P)$. This contradicts the fact that $|N_G(Q)| = 3 \cdot 17$.

We can refine this method by not requiring P and Q to be of prime order. Namely, if p and q are distinct primes dividing $|G|$ such that $Q \in Syl_q(G)$ and $p \mid |N_G(Q)|$, let $P \in Syl_p(N_G(Q))$. We can then apply Sylow's Theorems in $N_G(Q)$ to see whether $P \trianglelefteq N_G(Q)$, and if so, force $N_G(P)$ to be of small index. If P is a Sylow p-subgroup of the whole group G, we can use the congruence part of Sylow's Theorem to put further restrictions on $|N_G(P)|$ (as we did in the preceding example). If P is not a Sylow p-subgroup of G, then by the second part of Sylow's Theorem $P \leq P^* \in Syl_p(G)$. In this case since $P < P^*$, Theorem 1(4) shows that $P < N_{P^*}(P)$. Thus $N_G(P)$ (which contains $N_{P^*}(P)$) has order divisible by a larger power of p than divides $|P|$ (as well as being divisible by $|Q|$).

For example, there are no simple groups of order 3675. If there were, let G be a simple group of order $3675 = 3 \cdot 5^2 \cdot 7^2$. The only possibility for n_7 is 15, so for $Q \in Syl_7(G)$, $|G : N_G(Q)| = 15$ and $|N_G(Q)| = 245 = 5 \cdot 7^2$. Let $N = N_G(Q)$ and let $P \in Syl_5(N)$. By the congruence conditions of Sylow's Theorem applied in N we get $P \trianglelefteq N$. Since $|P| = 5$, P is not itself a Sylow 5-subgroup of G so P is contained in some Sylow 5-subgroup P^* of G. Since P is of index 5 in the 5-group P^*, $P \trianglelefteq P^*$ by Theorem 1, that is $P^* \leq N_G(P)$. This proves:

$$\langle N, P^* \rangle \leq N_G(P) \quad \text{so} \quad 7^2 \cdot 5^2 \mid |N_G(P)|.$$

Thus $|G : N_G(P)| \mid 3$, which is impossible since P is not normal and G has no subgroup of index 3.

Studing Normalizers of Intersections of Sylow p-Subgroups

One of the reasons the counting arguments in the first method above do not immediately generalize to Sylow subgroups which are not of prime order is because if $P \in Syl_p(G)$ for some prime p and $|P| = p^a$, $a \geq 2$, then it need not be the case that distinct conjugates of P intersect in the identity subgroup. If distinct conjugates of P *do* intersect in the identity, we can again count to find that the number of elements of p-power order is $n_p(|P| - 1)$.

Suppose, however, there exists $R \in Syl_p(G)$ with $R \neq P$ and $P \cap R \neq 1$. Let

$P_0 = P \cap R$. Then $P_0 < P$ and $P_0 < R$, hence by Theorem 1

$$P_0 < N_P(P_0) \quad \text{and} \quad P_0 < N_R(P_0).$$

One can try to use this to prove that the normalizer in G of P_0 is sufficiently large (i.e. of sufficiently small index) to obtain a contradiction by previous methods (note that this normalizer is a proper subgroup since $P_0 \neq 1$).

One special case where this works particularly well is when $|P_0| = p^{a-1}$ i.e. the two Sylow p-subgroups R and P have large intersection. In this case set $N = N_G(P_0)$. Then by the above reasoning (i.e. since P_0 is a maximal subgroup of the p-groups P and R), $P_0 \trianglelefteq P$ and $P_0 \trianglelefteq R$, that is,

$$N \text{ has 2 distinct Sylow } p\text{-subgroups: } P \text{ and } R.$$

In particular, $|N| = p^a k$, where (by Sylow's Theorem) $k \geq p + 1$.

Recapitulating, if Sylow p-subgroups pairwise intersect in the identity, then counting elements of p-power order is possible; otherwise there is some intersection of Sylow p-subgroups whose normalizer is "large". Since for an arbitrary group order one cannot necessarily tell which of these two phenomena occurs, it may be necessary to split the non-simplicity argument into two (mutually exclusive) cases and derive a contradiction in each. This process is especially amenable when the order of a Sylow p-subgroup is p^2 (for example, this line of reasoning was used to count elements of 2-power order in the proof that a simple group of order 60 is isomorphic to A_5 — Proposition 23, Section 4.5).

Before proceeding with an example we state a lemma which gives a sufficient condition to force a nontrivial Sylow intersection:

Lemma 13. In a finite group G if $n_p \not\equiv 1 (\text{mod } p^2)$, then there are distinct Sylow p-subgroups P and R of G such that $P \cap R$ is of index p in both P and R (hence is normal in each).

Proof: The argument is an easy refinement of the proof of the congruence part of Sylow's Theorem (cf. the exercises at the end of Section 4.5). Let P act by conjugation on the set $Syl_p(G)$. Let $\mathcal{O}_1, \ldots, \mathcal{O}_s$ be the orbits under this action with $\mathcal{O}_1 = \{P\}$. If p^2 divides $|P : P \cap R|$ for all Sylow p-subgroups R of G different from P, then each $\mathcal{O}_i$ has size divisible by p^2, $i = 2, 3, \ldots, s$. In this case, since n_p is the sum of the lengths of the orbits we would have $n_p = 1 + kp^2$, contrary to assumption. Thus for some $R \in Syl_p(G)$, $|P : P \cap R| = p$.

For example, there are no simple groups of order 3159. If there were, let G be a simple group of order $3159 = 3^4 \cdot 13$ and let $P \in Syl_3(G)$. We must have $n_3 = 13$. But $13 \not\equiv 1 (\text{mod } 3^2)$ so there exist $P, R \in Syl_3(G)$ such that $|P \cap R| = 3^3$. Let $N = N_G(P \cap R)$, so by the above arguments $P, R \leq N$. Thus $3^4 \mid |N|$ and $|N| > 3^4$. The only possibility is $N = G$, i.e. $P \cap R \trianglelefteq G$, a contradiction.

Simple Groups of Order 168

We now show how many of our techniques can be used to unravel the structure of a simple group (even though the subgroup structure of the normal subgroups is triv-

ial). We assume that there is a simple group of order 168 and work out many of its properties: the number and structure of its Sylow subgroups, the possible orders of elements etc.). All of these calculations are based only on the order and the simplicity of the group. Following this, we prove that in fact there is a simple group of order 168: $GL_2(\mathbb{F}_3)$. Throughout this discussion G is assumed to be a simple group of order $168 = 2^3 \cdot 3 \cdot 7$.

(1) $n_7 = 8$ and G has 48 elements of order 7.

The only possibilities allowed by Sylow's Theorem are $n_7 = 1$ and $n_7 = 8$. The latter occurs by the simplicity of G. Now G has 8 Sylow 7-subgroups which pairwise intersect in the identity. Each Sylow 7-subgroup contains 6 elements of order 7. Thus G has $8 \cdot 6 = 48$ elements of order 7.

(2) $n_3 = 7$ or 28.

To see this we first prove that if $Q \in Syl_3(G)$, then 7 does not divide $|N_G(Q)|$, i.e. $7 \mid n_3$. For if 7 divided $|N_G(Q)|$ let $P \in Syl_7(N_G(Q))$. Then PQ is a group of order 21 with a normal Sylow 3-subgroup. By Sylow's Theorem its Sylow 7-subgroup is also normal, hence PQ is abelian (hence cyclic). It follows from (1) that G is isomorphic to a subgroup of S_8. This leads to a contradiction because S_8 has no elements of order 21 (check cycle types). This contradiction proves that $7 \mid n_3$. The only permissible values of n_3 which are divisible by 7 are 7 and 28, which is (2).

(3) the minimal possible index of a proper subgroup is 7 and G is isomorphic to a subgroup of A_7.

The first assertion is immediate from the fact that 168 does not divide 6!. To establish the second we must prove G has a subgroup of index 7, for then G is isomorphic to a subgroup of S_7 and hence, by Proposition 12, to a subgroup of A_7. If $n_3 = 7$, we are done, so assume $n_3 = 28$. Since the minimal index of a proper subgroup is 7, the only possibilities for n_2 are 7 and 21. In the former case we are again done, so assume $n_2 = 21$.

If every pair of distinct Sylow 2-subgroups intersect in the identity, counting elements shows that there are 147 nonidentity elements of 2-power order in G. This is impossible because G has 48 elements of order 7. There must therefore be distinct Sylow 2-subgroups T_1 and T_2 which intersect nontrivially. Over all pairs of distinct Sylow 2-subgroups let T_1 and T_2 be chosen to have intersection of largest possible order. Let $S = T_1 \cap T_2$ and let $N = N_G(S)$. By Theorem 1(4)

$$1 < S < N_{T_1}(S) \leq N \qquad \text{and} \qquad 1 < S < N_{T_2}(S) \leq N \qquad (*)$$

hence $4 \mid |N|$. Since the minimal index of a proper subgroup is 7, the only possibilities for $|N|$ are: 4, 8, 12, 24 or 28. By eliminating all other numbers in this list we prove $|N| = 24$; then $|G : N| = 7$, so G has a subgroup of index 7.

If $|N| = 4$, we must have $N = N_{T_1}(S) = N_{T_2}(S)$. In this case $N \leq T_1 \cap T_2 = S$ contrary to $(*)$.

If $|N| = 8$ then $N \in Syl_2(G)$. Since T_1 and T_2 are distinct, at least one of

these must be different from N. Choose notation so that $T_1 \neq N$. Now T_1 and N are distinct Sylow 2-subgroups of G which contain S. Now $N_{T_1}(S) \leq N_G(S) = N$ and, by $(*)$, $S < N_{T_1}(S) \leq T_1$. Thus N and T_1 are distinct Sylow 2-subgroups of G whose intersection contains the group $N_{T_1}(S)$ which has order larger than $|S|$. This contradicts the choice of S and so proves that N cannot have order 8.

If $|N| = 12$, by the classification of groups of order 12 (Section 4.5) N has either a normal Sylow 2-subgroup or a normal Sylow 3-subgroup. Since we are in the situation where $n_3 = 28$ the normalizer of a Sylow 3-subgroup has order 6, so the latter possibility is ruled out. If, however, N has a normal (hence unique) Sylow 2-subgroup, it follows from $(*)$ that $N_{T_1}(S) = N_{T_2}(S)$, contrary to $T_1 \cap T_2 = S$. This proves $|N| \neq 12$.

Finally, if $|N| = 28$, by Sylow's Theorem N has a normal Sylow 7-subgroup. This is impossible because (by (1)) the normalizer of a Sylow 7-subgroup of G has order 21. This contradiction completes the proof of (3).

(4) $n_2 = 21$ and Sylow 2-subgroups of G are isomorphic to D_8.

By (3) we may consider G as a subgroup of A_7. By order considerations Sylow 2-subgroups of G and A_7 are isomorphic. But A_7 has a subgroup isomorphic to D_8 (namely, the usual symmetries of a square with vertices labelled 1,...,4 but with each reflection multiplied by the transposition (6 7) to make it an even permutation). Thus Sylow 2-subgroups of both A_7 and G are dihedral. To see that $n_2 = 21$ it is necessary only to eliminate the possibility $n_2 = 7$. If $n_2 = 7$, let $T \in Syl_2(G)$ and let $N = N_G(T)$, a group of order 24. Since $T \cong D_8$ has a characteristic subgroup isomorphic to Z_4, N has a normal cyclic subgroup of order 4. Since (cf. Section 4.4) $\mathrm{Aut}(Z_4) \cong Z_2$, any element of order 3 in N must centralize this subgroup. Thus N contains commuting elements of orders 3 and 4, hence contains an element of order 12. This is impossible since A_7 contains no elements of order 12 (although S_7 does – check the cycle types). This contradiction proves that $n_2 \neq 7$, hence $n_2 = 21$ is the only remaining possibility.

(5) G has no element of order 6.

We first show that the elements of G of order 3 are products of two 3-cycles in A_7. To see this note that by Proposition 12(2) and by counting arguments as used in the discussion on permutation representations we have

$$|N_{A_7}(P)| = 21, \qquad \text{for } P \in Syl_7(A_7).$$

Thus the normalizer of a Sylow 7-subgroup of G and of A_7 coincide. By inspection, $P = \langle(1\,2\,3\,4\,5\,6\,7)\rangle \in Syl_7(A_7)$ and $(2\,3\,5)(4\,7\,6)$ is an element of order 3 normalizing P (recall how to conjugate in S_n – Section 4.3). Thus elements of order 3 normalizing Sylow 7-subgroups are products of two 3-cycles. Since G has, by Sylow's Theorem, only one conjugacy class of subgroups of order 3 it follows that *every* element of order 3 in G is a product of two 3-cycles. Now one easily computes (as we did in Section 4.3) that no permutation in A_7 of order 2 commutes with a product of two 3-cycles, so G cannot have an element of order 6.

(6) $n_3 = 28$ and G has 56 elements of order 3.

If $n_3 \neq 28$ then by (2) we must have $n_3 = 7$. Let Q be a Sylow 3-subgroup of G and let T be a Sylow 2-subgroup of $N_G(Q)$, so $|T| = 8$. Since $N_G(Q)/C_G(Q)$ is isomorphic to a subgroup of $\mathrm{Aut}(Q)$ and the latter group has order 2 (see Section 4.4), we must have $|T/(T \cap C_G(Q))| \leq 2$. This forces $T \cap C_G(Q)$ to have order at least 4, so T contains an element of order 2 commuting with an element of order 3 in Q. This contradicts the fact that G does not have any elements of order 6 and so proves that $n_3 = 28$. Each Sylow 3-subgroup contains 2 elements of order 3, so there are $28 \cdot 2 = 56$ elements of order 3.

It is now straightforward to compute all the conjugacy classes of G and their sizes. Let x be an element of order 7 in G. Since an element of order 7 in S_7 (i.e. a 7-cycle) commutes only with its own powers in S_7 the same must be true in G, i.e. $C_G(x) = \langle x \rangle$. Thus the number of conjugates of x is $|G : C_G(x)| = 24$. By (1) there must therefore be 2 conjugacy classes of elements of order 7. Moreover, since $|N_G(\langle x \rangle)|$ is odd it follows that x and x^{-1} are not conjugate in G, and so these represent the two conjugacy classes of elements of order 7.

Since G has no elements of order 6 or 21, an element y of order 3 in G commutes only with its powers: $C_G(y) = \langle y \rangle$. Thus y has $|G : C_G(y)| = 56$ conjugates and so all elements of order 3 in G are conjugate.

A Sylow 2-subgroup T of G is dihedral and so contains two elements of order 4, u and u^{-1}; these are conjugate in T. By Sylow's Theorem every element of order 4 is conjugate to some element (of order 4) in T, hence G has one conjugacy class of elements of order 4. Since G has no elements of order 12 or 28 and has no abelian subgroup of order 8, it follows that $C_G(u) = \langle u \rangle$. Thus u has $|G : C_G(u)| = 42$ conjugates.

Finally, let z be the element of order 2 in the center of the Sylow 2-subgroup T. Since G has no elements of order 6 or 14, $C_G(z) = T$. Thus z has $|G : C_G(z)| = 21$ conjugates. The conjugates of 1, x, x^{-1}, y, u and z account for all 168 elements. In particular, all elements of order 2 in G are conjugate to z.

We now summarize the results we have just proved as parts (1) to (4) of Proposition 14. Parts (5) and (6) may be established by the reader by similar arguments, building on the preceding facts.

Proposition 14. If G is a simple group of order 168, then the following hold:

(1) $n_2 = 21$, $n_3 = 7$ and $n_7 = 8$

(2) Sylow 2-subgroups of G are dihedral, Sylow 3-subgroups and Sylow 7-subgroups are cyclic

(3) G is isomorphic to a subgroup of A_7 and G has no subgroup of index ≤ 6

(4) the conjugacy classes of G are: the identity; two classes of elements of order 7 each of which contains 24 elements (represented by any element of order 7 and its inverse); one class of elements of order 3 containing 56 elements; one class of elements of order 4 containing 42 elements; one class of elements of order 2 containing 21 elements (in particular, every element of G has order a power of a prime)

(5) every maximal subgroup has index either 7 or 8

(6) G has three conjugacy classes of maximal subgroups, two of which are isomorphic to S_4 and one of which is isomorphic to the non-abelian group of order 21.

All of the calculations above were predicated on the assumption that there exists a simple group of order 168. The fact that none of these arguments leads to a contradiction does not *prove* the existence of such a group, but rather just gives strong evidence that there *may* be a simple group of this order. We now produce such a group.

Theorem 15. $GL_3(\mathbb{F}_2)$ is a simple group of order 168.

Proof: Let $G = GL_3(\mathbb{F}_2)$ where this group can be described either as the automorphism group of the elementary abelian 2-group $Z_2 \times Z_2 \times Z_2$ (cf. Section 4.4) or as the group of 3×3 matrices with entries from $\mathbb{F}_2$ and nonzero determinant. The formula for the orders of the finite linear groups listed in Section 1.4 will be proved in Chapter 11; we use this formula to obtain that $|G| = 168$. These are the only facts we shall use about G.

Since G permutes the 7 nonidentity elements of $Z_2 \times Z_2 \times Z_2$ with only the identity element of G fixing everything, G is isomorphic to a subgroup of S_7. We henceforth identify G with its image in S_7. Let $P \in Syl_7(G)$, so $P \in Syl_7(S_7)$ as well. As before, the normalizer in S_7 of P has order 42, so $|N_G(P)| \mid 42$. By Sylow's Theorem the only possibilities for n_7 are 1 or 8. The former is impossible because then we would have $168 = |G| = |N_G(P)|$ dividing 42. Thus $n_7 = 8$ and $|N_G(P)| = 21$.

We show that G acts as even permutations in this degree 7 permutation representation. For if not, then $G \cap A_7$ is a subgroup of index 2 in G. It follows from Sylow's Theorem that this subgroup of order 84 has a normal Sylow 7-subgroup, contrary to the fact that the normalizer in G of a Sylow 7-subgroup has order 21. This proves $G \leq A_7$. From this we obtain that Sylow 2-subgroups of G are dihedral of order 8 (recall that A_7 contains a subgroup isomorphic to D_8, hence Sylow 2-subgroups of A_7 are of this type).

Assume now by way of contradiction that G has a proper normal subgroup H and consider first the case when 7 divides the order of H. Let P be a Sylow 7-subgroup of H. By Frattini's Argument (Proposition 6)

$$G = HN_G(P) \quad \text{and} \quad |G : H| \text{ divides } |N_G(P)|.$$

Since $|N_G(P)| = 21$ and $7 \mid |H|$, we must have $|G : H| = 3$, that is, $|H| = 56$. If H had a normal (hence characteristic) Sylow 7-subgroup, then so would G, a contradiction. Thus H has more than one Sylow 7-subgroup. By counting elements of order 7 in the group H of order 56 we see that there are only 8 elements in H which are not of order 7. Thus H has a normal Sylow 2-subgroup — call it D. Since $D \cong D_8$, and $Z(D)$ char $D \trianglelefteq H$, we get that $Z(D)$ is a normal subgroup of H of order 2. It follows that $Z(D) \leq Z(H)$. But then $Z(D)$ commutes with P, contrary to no element of order 7 in S_7 commuting with any element of order 2 (a 7-cycle in S_7 commutes only with its powers). This proves that 7 cannot divide the order of any proper normal subgroup of G.

If 3 divides $|H|$, let Q be a Sylow 3-subgroup of H. By Frattini's Argument 7 divides $|N_G(Q)|$. It follows that $N_G(Q)$ has an abelian (hence cyclic) subgroup of order 21, contrary again to an element of order 7 in S_7 commuting only with its powers.

We have reduced to the case when $|H| = 2$, 4 or 8. It follows from the congruence part of Sylow's Theorem applied to the group G/H (of order 84, 42 or 21 respectively) that G/H has a (proper) normal Sylow 7-subgroup. If we let H_1 be the preimage in G of this subgroup, then H_1 is normal in G and $|H_1| = 7 \cdot |H| < |G|$. This contradicts an earlier argument which showed that G has no proper normal subgroup whose order is divisible by 7. The proof of the theorem is complete.

With a certain amount of elementary argument (calculations in A_7, for example see Exercise 27) it can be shown that there is, up to isomorphism, a *unique* simple group of order 168 (it is not always the case that there is at most one simple group of a given order: there are 2 nonisomorphic simple groups of order $\frac{1}{2}8!$.)

Remarks on the Existence Problem for Groups

As in other areas of mathematics (such as the theory of differential equations) one may hypothesize the existence of a mathematical system (e.g. solution to an equation) and derive a great deal of information about this proposed system. In general, if after considerable effort no contradiction is reached based on the initial hypotheses one begins to suspect that there actually is a system which does satisfy the conditions hypothesized. However, no amount of consistent data will *prove* existence. Suppose we carried out an analysis of a hypothetical simple group G of order $3^3 \cdot 7 \cdot 13 \cdot 409$ analogous to our analysis of a simple group of order 168 (which we showed to exist). After a certain amount of effort we could show that there are unique possible Sylow numbers:

$$n_3 = 7 \cdot 409 \qquad n_7 = 3^2 \cdot 13 \cdot 409 \qquad n_{13} = 3^2 \cdot 7 \cdot 409 \qquad n_{409} = 3^2 \cdot 7 \cdot 13.$$

We could further show that such a G would have no elements of order pq, p and q distinct primes, no elements of order 9, and that distinct Sylow subgroups would intersect in the identity. We could then count the elements in Sylow p-subgroups for all primes p and we would find that these would total to exactly $|G|$. At this point we would have the complete subgroup structure and class equation for G. We might then guess that there *is* a simple group of this order, but the Feit–Thompson Theorem asserts that there are no simple groups of odd composite order. The point is that even though we have precisely as much data in this case as we had in the order 168 situation, we cannot prove existence by these techniques. Proving existence requires techniques for *constructing* groups.

When we are dealing with non-simple groups we have at least one method of building larger groups from smaller ones: semidirect products. Even though this method is fairly restrictive it gives an inkling that non-simple groups might be built up from smaller groups in some constructive fashion. This process breaks down completely for simple groups. We thus achieve an appreciation of the division of tasks specified by the Hölder Program, namely the determination of the simple groups and then the determination of how these groups are put together to form larger groups.

The study of simple groups involves many of the same tools as the study of non-simple groups (to unravel their subgroup structures, etc.) but also requires other techniques for their construction. These often involve algebraic/geometric techniques which construct simple groups as automorphisms of interesting mathematical objects (sets, vector spaces, graphs, etc.). For example, the simple group A_n ($n \geq 5$) is a subgroup of index 2 in the group of all set automorphisms of a set of cardinality n (namely, the subgroup of all even permutations) and the simple group $GL_3(\mathbb{F}_2)$ is the group of automorphisms of a 3-dimensional vector space over the field of 2 elements. While we have come a long way in the analysis of finite groups, there are a number of different areas in this branch of mathematics on which we have just touched.

The analysis of infinite groups generally involves quite different techniques and in the next section we introduce some basic infinite group theory.

EXERCISES

Counting elements:

1. Prove that for fixed $P \in Syl_p(G)$ if $P \cap R = 1$ for all $R \in Syl_p(G) - \{P\}$, then $P_1 \cap P_2 = 1$ whenever P_1 and P_2 are distinct Sylow p-subgroups of G. Deduce in this case that the number of nonidentity elements of p-power order in G is $(|P| - 1)|G : N_G(P)|$.
2. In the group $S_3 \times S_3$ exhibit a pair of Sylow 2-subgroups that intersect in the identity and exhibit another pair that intersect in a group of order 2.
3. Prove that if $|G| = 380$ then G is not simple. [Just count elements of odd prime order.]
4. Prove that there are no simple groups of order 80, 1053, 3875 or 5313.
5. Let G be a solvable group of order pm, where p is a prime not dividing m, and let $P \in Syl_p(G)$. If $N_G(P) = P$, prove that G has a normal subgroup of order m. Where was the solvability of G needed in the proof? (This result is true for nonsolvable groups as well — it is a special case of *Burnside's N/C-Theorem*.)

Exploiting subgroups of small index:

6. Prove that there are no simple groups of order 2205, 4125, 5103, 6545 or 6435.

Permutation representations:

7. Prove that there are no simple groups of order 1755 or 5265. [Use Sylow 3-subgroups to show $G \leq S_{13}$ and look at the normalizer of a Sylow 13-subgroup.]
8. Prove that there are no simple groups of order 792 or 918.
9. Prove that there are no simple groups of order 336.

Playing p-subgroups off against each other:

10. Prove that there are no simple groups of order 4095, 4389, 5313 or 6669.
11. Prove that there are no simple groups of order 4851 or 5145.
12. Prove that there are no simple groups of order 9555. [Let $Q \in Syl_{13}(G)$ and let $P \in Syl_7(N_G(Q))$. Argue that $Q \trianglelefteq N_G(P)$ — why is this a contradiction?]

Normalizers of Sylow intersections:

13. Prove that there are no simple groups of order 144, 351, 525 or 2025.

General exercises:

14. Classify groups of order 105.
15. Prove that there are no non-abelian simple groups of odd order < 10000.
16. Prove that there is no simple group of order 420.

17. Prove that there are no simple groups of even order < 500 except for orders 2, 60, 168 and 360.
18. Prove that if G is a group of order 36 then G has either a normal Sylow 2-subgroup or a normal Sylow 3-subgroup.
19. Show that a group of order 12 with no subgroup of order 6 is isomorphic to A_4.
20. Show that a group of order 24 with no element of order 6 is isomorphic to S_4.
21. Generalize Lemma 13 by proving that if $n_p \not\equiv 1 (\text{mod } p^k)$ then there are distinct Sylow p-subgroups P and R of G such that $P \cap R$ is of index $\leq p^{k-1}$ in both P and R.
22. Suppose over all pairs of distinct Sylow p-subgroups of G, P and R are chosen with $|P \cap R|$ maximal. Prove that $N_G(P \cap R)$ is not a p-group.
23. Let A and B be normal subsets of a Sylow p-subgroup P of G. Prove that if A and B are conjugate in G then they are conjugate in $N_G(P)$.
24. Let G be a group of order pqr where p, q and r are primes with $p < q < r$. Prove that a Sylow r-subgroup of G is normal.
25. Let G be a simple group of order p^2qr where p, q and r are primes. Prove that $|G| = 60$.
26. Prove or construct a counterexample to the assertion: if G is a group of order 168 with more than one Sylow 7-subgroup then G is simple.
27. Prove that there is a unique simple group of order 168. [Any such is contained in A_7 and, up to conjugacy in A_7, contains the normalizer, N, of the Sylow 7-subgroup described in point (5) in the proof of Proposition 14. Fix an element of order 3 in N and show that this element (which is a product of two 3-cycles) is contained in exactly three subgroups of order 6 in A_7 — call these three subgroups H_1, H_2, and H_3. Deduce that there are at most three conjugacy classes of simple subgroups of order 168 in A_7, namely $\langle N, H_i \rangle, i = 1, 2, 3$. Show that any subgroup of A_7 of order 168 has exactly 15 conjugates in A_7 and deduce that A_7 contains either 15, 30 or 45 simple subgroups of order 168. Prove that the number of simple subgroups of A_7 of order 168 cannot be odd by letting a transposition permute them by conjugation and using Exercise 4 in Section 3.5. By the same reasoning show that if A_7 has 30 simple subgroups of order 168 then these subgroups are all conjugate in S_7, hence are all isomorphic. This reasoning proves that exactly two of the subgroups $\langle N, H_i \rangle$ are simple and of order 168. Taking a few products of elements will quickly show which of the three is not a simple group of order 168 — in fact, this subgroup is all of A_7.]
28. Let G be a simple group of order $3^3 \cdot 7 \cdot 13 \cdot 409$. Compute all permissible values of n_p for each $p \in \{3, 7, 13, 409\}$ and reduce to the case where there is a unique possible value for each n_p.
29. Given the information on the Sylow numbers for a hypothetical simple group of order $3^3 \cdot 7 \cdot 13 \cdot 409$, prove that there is no such group. [Work with the permutation representation of degree 1218.]
30. Suppose G is a simple group of order 720. Find as many properties of G as you can (Sylow numbers, isomorphism type of Sylow subgroups, conjugacy classes, etc.). Is there such a group?

6.3 A WORD ON FREE GROUPS

In this section we introduce the basic theory of so-called free groups. This will enable us to make precise the notions of generators and relations which were used in earlier chapters. The results of this section rely only on the basic theory of homomorphisms.

The basic idea of a free group $F(S)$ generated by a set S is that there are no relations satisfied by any of the elements in S (S is "free" of relations). For example, if S is the set $\{a, b\}$ then the elements of the free group on the two generators a and b are of the form a, aa, ab, $abab$, bab, etc., called *words* in a and b, together with the inverses of these elements, and all these elements are considered distinct. If we group like terms together, then we obtain elements of the familiar form a, b^{-3}, $aba^{-1}b^2$ etc. Such elements are multiplied by concatenating their words (for example, the product of aba and $b^{-1}a^3b$ would simply be $abab^{-1}a^3b$). It is natural at the outset (even before we know S is contained in some group) to simply *define* $F(S)$ to be the set of all words in S, where two such expressions are multiplied in $F(S)$ by concatenating them. Although in essence this is what we do, it is necessary to be more formal in order to prove that this concatenation operation is well defined and associative. After all, even the familiar notation a^n for the product $a \cdot a \cdots a$ (n terms) is permissible only because we know that this product is independent of the way it is bracketed (cf. the generalized associative law in Section 1.1). The formal construction of $F(S)$ is carried out below for an arbitrary set S.

One important property reflecting the fact that there are no relations that must be satisfied by the generators in S is that any *map* from the *set* S to a group G can be uniquely extended to a *homomorphism* from the *group* $F(S)$ to G (basically since we have specified where the generators must go and the images of all the other elements are uniquely determined by the homomorphism property — the fact that there are no relations to worry about means that we can specify the images of the generators *arbitrarily*). This is frequently referred to as the *universal* property of the free group and in fact characterizes the group $F(S)$.

The notion of "freeness" occurs in many algebraic systems and it may already be familiar (using a different terminology) from elementary vector space theory. When the algebraic systems are vector spaces, $F(S)$ is simply the vector space which has S as a basis. Every vector in this space is a unique linear combination of the elements of S (the analogue of a "word"). Any set map from the basis S to another vector space V extends uniquely to a linear transformation (i.e. vector space homomorphism) from $F(S)$ to V.

Before beginning the construction of $F(S)$ we mention that one often sees the universal property described in the language of commutative diagrams. In this form it reads (for groups) as follows: given any set map φ from the set S to a group G there is a unique homomorphism $\Phi : F(S) \to G$ such that $\Phi|_S = \varphi$ i.e. such that the following diagram commutes:

$$\begin{tikzcd} S \arrow[r, "\text{inclusion}"] \arrow[dr, "\varphi"'] & F(S) \arrow[d, "\Phi"] \\ & G \end{tikzcd}$$

As mentioned above, the only difficulty with the construction of $F(S)$ is the verification that the concatenation operation on the words in $F(S)$ is well defined and associative. To prove the associative property for multiplication of words we return to the most basic level where all the exponents in the words of S are ± 1.

We first introduce inverses for elements of S and an identity.

Let S^{-1} be any set disjoint from S such that there is a bijection from S to S^{-1}. For each $s \in S$ denote its corresponding element in S^{-1} by s^{-1} and similarly for each $t \in S^{-1}$ let the corresponding element of S be denoted by t^{-1} (so $(s^{-1})^{-1} = s$). Take a singleton set not contained in $S \cup S^{-1}$ and call it $\{1\}$. Let $1^{-1} = 1$ and for any $x \in S \cup S^{-1} \cup \{1\}$ let $x^1 = x$.

Next we describe the elements of the free group on the set S. A *word* on S is by definition a sequence

$$(s_1, s_2, s_3, \dots) \quad \text{where } s_i \in S \cup S^{-1} \cup \{1\} \text{ and } s_i = 1 \text{ for all } i \text{ sufficiently large}$$

(that is, for each sequence there is an N such that $s_i = 1$ for all $i \geq N$). Thus we can think of a word as a finite product of elements of S and their inverses (where repetitions are allowed). Next, in order to assure uniqueness of expressions we consider only words which have no obvious "cancellations" between adjacent terms (such as $baa^{-1}b = bb$). The word $(s_1, s_2, s_3, \dots)$ is said to be *reduced* if

(1) $s_{i+1} \neq s_i^{-1}$ for all i with $s_i \neq 1$, and
(2) if $s_k = 1$ for some k, then $s_i = 1$ for all $i \geq k$.

The reduced word $(1, 1, 1, \dots)$ is called the *empty word* and is denoted by 1. We now simplify the notation by writing the reduced word $(s_1^{\epsilon_1}, s_2^{\epsilon_2}, \dots, s_n^{\epsilon_n}, 1, 1, 1, \dots)$, $s_i \in S$, $\epsilon_i = \pm 1$, as $s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}$. Note that by definition, reduced words $r_1^{\delta_1} r_2^{\delta_2} \dots r_m^{\delta_m}$ and $s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}$ are equal if and only if $n = m$ and $\delta_i = \epsilon_i$, $1 \leq i \leq n$. Let $F(S)$ be the set of reduced words on S and embed S into $F(S)$ by

$$s \mapsto (s, 1, 1, 1, \dots).$$

Under this set injection we identify S with its image and henceforth consider S as a subset of $F(S)$. Note that if $S = \emptyset$, $F(S) = \{1\}$.

We are now in a position to introduce the binary operation on $F(S)$. The principal technical difficulty is to ensure that the product of two reduced words is again a *reduced* word. Although the definition appears to be complicated it is simply the formal rule for "successive cancellation" of juxtaposed terms which are inverses of each other (e.g. $ab^{-1}a$ times $a^{-1}ba$ should reduce to aa). Let $r_1^{\delta_1} r_2^{\delta_2} \dots r_m^{\delta_m}$ and $s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}$ be reduced words and assume first that $m \leq n$. Let k be the smallest integer in the range $1 \leq k \leq m + 1$ such that $s_k^{\epsilon_k} \neq r_{m-k+1}^{-\delta_{m-k+1}}$. Then the product of these reduced words is defined to be:

$$(r_1^{\delta_1} r_2^{\delta_2} \dots r_m^{\delta_m})(s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}) = \begin{cases} r_1^{\delta_1} \dots r_{m-k+1}^{\delta_{m-k+1}} s_k^{\epsilon_k} \dots s_n^{\epsilon_n}, & \text{if } k \leq m \\ s_{m+1}^{\epsilon_{m+1}} \dots s_n^{\epsilon_n}, & \text{if } k = m + 1 \leq n \\ 1, & \text{if } k = m + 1 \text{ and } m = n \end{cases}$$

The product is defined similarly when $m \geq n$, so in either case it results in a reduced word.

Theorem 15. $F(S)$ is a group under the binary operation defined above.

Proof: One easily checks that 1 is an identity and that the inverse of the reduced word $s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}$ is the reduced word $s_n^{-\epsilon_n} s_{n-1}^{-\epsilon_{n-1}} \dots s_1^{-\epsilon_1}$. The difficult part of the proof is the verification of the associative law. This can be done by induction on the "length" of the words involved and considering various cases or one can proceed as follows: For each $s \in S \cup S^{-1} \cup \{1\}$ define $\sigma_s : F(S) \to F(S)$ by

$$\sigma_s(s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}) = \begin{cases} s \cdot s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}, & \text{if } s_1^{\epsilon_1} \neq s^{-1} \\ s_2^{\epsilon_2} s_3^{\epsilon_3} \dots s_n^{\epsilon_n}, & \text{if } s_1^{\epsilon_1} = s^{-1}. \end{cases}$$

Since $\sigma_{s^{-1}} \circ \sigma_s$ is the identity map of $F(S) \to F(S)$, σ_s is a permutation of $F(S)$. Let $A(F)$ be the subgroup of the symmetric group on the set $F(S)$ which is generated by $\{\sigma_s \mid s \in S\}$. It is easy to see that the map

$$s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n} \mapsto \sigma_{s_1}^{\epsilon_1} \circ \sigma_{s_2}^{\epsilon_2} \circ \dots \circ \sigma_{s_n}^{\epsilon_n}$$

is a (set) bijection between $F(S)$ and $A(S)$ which respects their binary operations. Since $A(S)$ is a group, hence associative, so is $F(S)$.

The universal property of free groups now follows easily.

Theorem 16. Let G be a group, S a set and $\varphi : S \to G$ a set map. Then there is a unique group homomorphism $\Phi : F(S) \to G$ such that the following diagram commutes:

$$\begin{tikzcd} S \arrow[r, "\text{inclusion}"] \arrow[dr, "\varphi"'] & F(S) \arrow[d, "\Phi"] \\ & G \end{tikzcd}$$

Proof: Such a map Φ must satisfy $\Phi(s_1^{\epsilon_1} s_2^{\epsilon_2} \dots s_n^{\epsilon_n}) = \varphi(s_1)^{\epsilon_1} \varphi(s_2)^{\epsilon_2} \dots \varphi(s_n)^{\epsilon_n}$ if it is to be a homomorphism (which proves uniqueness), and it is straightforward to check that this map is in fact a homomorphism (which proves existence).

Corollary 17. $F(S)$ is unique up to a unique isomorphism which is the identity map on the set S.

Proof: This follows from the universal property. Suppose $F(S)$ and $F'(S)$ are two free groups generated by S. Since S is contained in both $F(S)$ and $F'(S)$, we have natural injections $S \hookrightarrow F'(S)$ and $S \hookrightarrow F(S)$. By the universal property in the theorem, it follows that we have unique associated group homomorphisms $\Phi : F(S) \to F'(S)$ and $\Phi' : F'(S) \to F(S)$ which are the both the identity on S. The composite $\Phi'\Phi$ is a homomorphism from $F(S)$ to $F(S)$ which is the identity on S, so by the uniqueness statement in the theorem, it must be the identity map. Similarly $\Phi\Phi'$ is the identity, so Φ is an isomorphism (with inverse Φ'), which proves the corollary.

Definition. The group $F(S)$ is called the *free group* on the set S. A group F is a *free group* if there is some set S such that $F = F(S)$ – in this case we call S a set of

free generators (or a *free basis*) of F. The cardinality of S is called the *rank* of the free group.

One can now simplify expressions in a free group by using exponential notation, so we write a^3b^{-2} instead of the formal reduced word $aaab^{-1}b^{-1}$. Expressions like aba, however, cannot be simplified in the free group on $\{a, b\}$. We mention one important theorem in this area:

Theorem 18. (Schreier) Subgroups of a free group are free.

This is not trivial to prove and we do not include a proof. There is a nice proof of this result using covering spaces — cf. *Trees* by J.-P. Serre, Springer-Verlag, 1980.

Presentations

Let G be any group. Then G is a homomorphic image of a free group: take $S = G$ and φ as the identity map from G to G, then Theorem 16 produces a (surjective) homomorphism from $F(G)$ onto G. More generally, if S is any subset of G such that $G = \langle S \rangle$, then again there is a unique surjective homomorphism from $F(S)$ onto G which is the identity on S. (Note that we can now independently formulate the notion that a subset *generates* a group by noting that $G = \langle S \rangle$ if and only if the map $\pi : F(S) \to G$ which extends the identity map of S to G is surjective.)

Definition. Let S be a subset of a group G such that $G = \langle S \rangle$.

(1) A *presentation* for G is a pair (S, R), where R is a set of words in $F(S)$ such that the normal closure of $\langle R \rangle$ in $F(S)$ (the smallest normal subgroup containing $\langle R \rangle$) equals the kernel of the homomorphism $\pi : F(S) \to G$ (where π extends the identity map from S to S). The elements of S are called *generators* and those of R are called *relations* of G.

(2) We say G is *finitely generated* if there is a presentation (S, R) such that S is a finite set and we say G is *finitely presented* if there is a presentation (S, R) with both S and R finite sets.

Note that if (S, R) is a presentation, the kernel of the map $F(S) \to G$ is not $\langle R \rangle$ itself but rather the (much larger) group generated by R and *all conjugates* of elements in R. Note that even for a fixed set S a group will have many different presentations (we can always throw redundant relations into R, for example). If G is finitely presented with $S = \{s_1, s_2, \ldots, s_n\}$ and $R = \{w_1, w_2, \ldots, w_k\}$, we write (as we have in preceding chapters):

$$G = \langle s_1, s_2, \ldots, s_n \mid w_1 = w_2 = \cdots = w_k = 1 \rangle$$

and if w is the word $w_1 w_2^{-1}$, we shall write $w_1 = w_2$ instead of $w = 1$.

Examples

(1) Every finite group is finitely presented. To see this let $G = \{g_1, \ldots, g_n\}$ be a finite group. Let $S = G$ and let $\pi : F(S) \to G$ be the homomorphism extending the

identity map of S. Let R_0 be the set of words $g_i g_j g_k^{-1}$, where $i, j = 1, \ldots, n$ and $g_i g_j = g_k$ in G. Clearly $R_0 \leq \ker \pi$. If N is the normal closure of R_0 in $F(S)$ and $\widetilde{G} = F(S)/N$, then G is a homomorphic image of $\widetilde{G}$ (i.e. π factors through N). Moreover, the set of elements $\{\widetilde{g}_i \mid i = 1, \ldots, n\}$ is closed under multiplication. Since this set generates $\widetilde{G}$, it must equal $\widetilde{G}$. Thus $|\widetilde{G}| = |G|$ and so $N = \ker \pi$ and (S, R_0) is a presentation of G.

This illustrates a sufficient condition for (S, R) to be a presentation for a given finite group G:

(i) S must be a generating set for G, and

(ii) any group generated by S satisfying the relations in R must have order $\leq |G|$.

(2) Abelian groups can be presented easily. For instance:

$$\mathbb{Z} \cong F(\{a\}) = \langle a \rangle,$$
$$\mathbb{Z} \times \mathbb{Z} \cong \langle a, b \mid [a, b] = 1 \rangle,$$
$$Z_n \times Z_m \cong \langle a, b \mid a^n = b^m = [a, b] = 1 \rangle.$$

(Recall: $[a, b] = a^{-1} b^{-1} a b$).

(3) Some familiar non-abelian groups introduced in earlier chapters have simple presentations:

$$D_{2n} = \langle r, s \mid r^n = s^2 = 1,\ s^{-1} r s = r^{-1} \rangle$$
$$Q_8 = \langle i, j \mid i^4 = 1,\ j^2 = i^2,\ j^{-1} i j = i^{-1} \rangle.$$

To check, for example, the presentation for D_{2n} note that the relations in $\langle r, s \mid r^n = s^2 = 1,\ s^{-1} r s = r^{-1} \rangle$ imply that this group has a normal subgroup (generated by r) of order $\leq n$ whose quotient is generated by s (which has order ≤ 2). Thus any group with these generators and relations has order at most $2n$. Since we already know of the existence of the group D_{2n} of order $2n$ satisfying these conditions, the abstract presentation must equal D_{2n}.

(4) In general it is extremely difficult even to determine if a given set of generators and relations is or is not the identity group (let alone determine whether it is some other nontrivial finite group). For example, in the following two presentations the first group is an *infinite* group and the second is the *identity* group (cf. *Trees*, Chapter 1):

$$\langle x_1, x_2, x_3, x_4 \mid x_2 x_1 x_2^{-1} = x_1^2,\ x_3 x_2 x_3^{-1} = x_2^2,\ x_4 x_3 x_4^{-1} = x_3^2,\ x_1 x_4 x_1^{-1} = x_4^2 \rangle$$
$$\langle x_1, x_2, x_3, \mid x_2 x_1 x_2^{-1} = x_1^2,\ x_3 x_2 x_3^{-1} = x_2^2,\ x_1 x_3 x_1^{-1} = x_3^2 \rangle.$$

(5) It is easy to see that S_n is generated by the transpositions $(1\,2), (2\,3), \ldots, (n-1\,n)$, and that these satisfy relations:

$$((i\,i{+}1)(i{+}1\,i{+}2))^3 = 1 \quad \text{and} \quad [(i\,i{+}1), (j\,j{+}1)] = 1, \text{ whenever } |i - j| \geq 2$$

(here $|i - j|$ denotes the absolute value of the integer $i - j$). One can prove by induction on n that these form a presentation of S_n:

$$S_n \cong \langle t_1, \ldots, t_{n-1} \mid t_i^2 = 1,\ (t_i t_{i+1})^3 = 1, \text{ and } [t_i, t_j] = 1 \text{ whenever } |i - j| \geq 2,\ 1 \leq i, j \leq n - 1 \rangle.$$

As mentioned in Section 1.6 we can use presentations of a group to find homomorphisms between groups or to find automorphisms of a group. We did this in classifying groups of order 6, for example, when we proved that any non-abelian group of order 6 was generated by an element of order 3 and an element of order 2 inverting it; thus there is a homomorphism from S_3 onto any non-abelian group of order 6 (hence an isomorphism, by computing orders). More generally, suppose G is presented by, say, generators a, b with relations $r_1, \ldots, r_k$. If a', b' are any elements of a group H satisfying these relations, there is a homomorphism from G into H. Namely, if $\pi : F(\{a, b\}) \to G$ is the presentation homomorphism, we can define $\pi' : F(\{a, b\}) \to H$ by $\pi'(a) = a'$ and $\pi'(b) = b'$. Then $\ker \pi \leq \ker \pi'$ so π' factors through $\ker \pi$ and we obtain:

$$G \cong F(\{a, b\})/\ker \pi \longrightarrow H.$$

In, particular, if $\langle a', b' \rangle = H = G$, this homomorphism is an automorphism of G. Conversely, any automorphism must send a set of generators to another set of generators satisfying the same relations. Thus, for example, $D_8 = \langle a, b \mid a^2 = b^4 = 1,\ aba = b^{-1} \rangle$ and any pair a', b' of elements, where a' is a noncentral element of order 2 and b' is of order 4, satisfies the same relations. Since there are four noncentral elements of order 2 and two elements of order 4, D_8 has 8 automorphisms.

Similarly, any pair of elements of order 4 in Q_8 which are not equal or inverses of each other necessarily generate Q_8 and satisfy the relations given in Example 3 above. It is easy to check that there are 24 such pairs, so

$$|\mathrm{Aut}(Q_8)| = 24.$$

Free objects can be constructed in (many, but not all) other categories. For instance, a *monoid* is a set together with a binary operation satisfying all of the group axioms except the axiom specifying the existence of inverses. Free objects in the category of monoids play a fundamental role in theoretical computer science where they model the behavior of machines (Turing machines, etc.). We shall encounter free algebras (i.e. polynomial algebras) and free modules in later chapters.

EXERCISES

1. Let F_1 and F_2 be free groups of finite rank. Prove that $F_1 \cong F_2$ if and only if they have the same rank. What facts do you need in order to extend your proof to infinite ranks (where the result is also true)?
2. Prove that if $|S| > 1$ then $F(S)$ is non-abelian.
3. Prove that the commutator subgroup of the free group on 2 generators is not finitely generated (in particular, subgroups of finitely generated groups need not be finitely generated).
4. Prove that every nonidentity element of a free group is of infinite order.
5. Establish a finite presentation for A_4 using 2 generators.
6. Establish a finite presentation for S_4 using 2 generators.
7. Prove that the following is a presentaion for the quaternion group of order 8:

$$Q_8 = \langle a, b \mid a^2 = b^2,\ a^{-1}ba = b^{-1} \rangle.$$

8. Use presentations to find the orders of the automorphism groups of the groups $Z_2 \times Z_4$ and $Z_4 \times Z_4$.

9. Prove that $\mathrm{Aut}(Q_8) \cong S_4$.

10. This exercise exhibits an automorphism of S_6 that is not inner (hence, together with Exercise 19 in Section 4.4 it shows that $|\mathrm{Aut}(S_6) : \mathrm{Inn}(S_6)| = 2$). Let $t_1' = (1\ 2)(3\ 4)(5\ 6)$, $t_2' = (1\ 4)(2\ 5)(3\ 6)$, $t_3' = (1\ 3)(2\ 4)(5\ 6)$, $t_4' = (1\ 2)(3\ 6)(4\ 5)$, and $t_5' = (1\ 4)(2\ 3)(5\ 6)$. Show that $t_1', \ldots, t_5'$ satisfy the following relations:
$(t_i')^2 = 1$ for all i,
$(t_i't_j')^2 = 1$ for all i and j with $|i - j| \geq 2$, and
$(t_i't_{i+1}')^3 = 1$ for all $i \in \{1, 2, 3, 4\}$.
Deduce that $S_6 = \langle t_1', \ldots, t_5' \rangle$ and that the map

$$(1\ 2) \mapsto t_1', \quad (2\ 3) \mapsto t_2', \quad (3\ 4) \mapsto t_3', \quad (4\ 5) \mapsto t_4', \quad (5\ 6) \mapsto t_5'$$

extends to an automorphism of S_6 (which is clearly not inner since it does not send transpositions to transpositions). [Use the presentation for S_6 described in Example 5.]

11. Let S be a set. The group with presentation (S, R), where $R = \{[s, t] \mid s, t \in S\}$ is called the *free abelian* group on S – denote it by $A(S)$. Prove that $A(S)$ has the following universal property: if G is any abelian group and $\varphi : S \to G$ is any set map, then there is a unique group homomorphism $\Phi : A(S) \to G$ such that $\Phi|_S = \varphi$. Deduce that if A is a free abelian group on a set of cardinality n then

$$A \cong \mathbb{Z} \times \mathbb{Z} \times \cdots \times \mathbb{Z} \quad (n \text{ factors}).$$

12. Let S be a set and let c be a positive integer. Formulate the notion of a *free nilpotent group* on S of nilpotence class c and prove it has the appropriate universal property with respect to nilpotent groups of class $\leq c$.

13. Prove that there cannot be a nilpotent group N generated by two elements with the property that *every* nilpotent group which is generated by two elements is a homomorphic image of N (i.e. the specification of the class c in the preceding problem was necessary).

Part II

RING THEORY

The theory of groups is concerned with general properties of certain objects having an algebraic structure defined by a single binary operation. The study of rings is concerned with objects possessing two binary operations (called addition and multiplication) related by the distributive laws. We first study analogues for the basic points of development in the structure theory of groups. In particular, we introduce subrings, quotient rings, ideals (which are the analogues of normal subgroups) and ring homomorphisms. We then focus on questions about general rings which arise naturally from the presence of two binary operations. Questions concerning multiplicative inverses lead to the notion of fields and eventually to the construction of some specific fields such as finite fields. The study of the arithmetic (divisibility, greatest common divisors, etc.) of rings such as the familiar ring of integers, $\mathbb{Z}$, leads to the notion of primes and unique factorizations in Chapter 8. The results of Chapters 7 and 8 are then applied to rings of polynomials in Chapter 9.

The basic theory of rings developed in Part II is the cornerstone for the remaining three parts of the book. The theory of ring actions (modules) comprises Part III of the book. There we shall see how the structure of rings is reflected in the structure of the objects on which they act and this will enable us to prove some powerful classification theorems. The structure theory of rings, in particular of polynomial rings, forms the basis in Part IV for the the theory of fields and polynomial equations over fields. There the rich interplay between ring theory, field theory and group theory leads to many beautiful results on the structure of fields and the theory of roots of polynomials. In Part V the study of certain specific kinds of rings (group rings) and the objects (modules) on which they act again gives deep classification theorems whose consequences are then exploited to provide new results and insights into finite groups.

7

Introduction to Rings

7.1 BASIC DEFINITIONS AND EXAMPLES

Definition

(1) A *ring* R is a set together with two binary operations $+$ and $\times$ (called addition and multiplication) satisfying the following axioms:

(i) $(R, +)$ is an *abelian* group,

(ii) $\times$ is associative : $(a \times b) \times c = a \times (b \times c)$ for all $a, b, c \in R$,

(iii) the *distributive laws* hold in R: for all $a, b, c \in R$

$$(a + b) \times c = (a \times c) + (b \times c) \quad \text{and} \quad a \times (b + c) = (a \times b) + (a \times c).$$

(2) The ring R is *commutative* if multiplication is commutative.

(3) The ring R is said to have an *identity* (or *contain a* 1) if there is an element $1 \in R$ with

$$1 \times a = a \times 1 = a \qquad \text{for all } a \in R.$$

We shall usually write simply ab rather than $a \times b$ for $a, b \in R$. The additive identity of R will always be denoted by 0 and the additive inverse of the ring element a will be denoted by $-a$.

The condition that R be a group under addition is a fairly natural one, but it may seem artificial to require that this group be *abelian*. One motivation for this is that if the ring R has a 1, the commutativity under addition is *forced* by the distributive laws. To see this, compute the product $(1 + 1)(a + b)$ in two different ways, using the distributive laws (but not assuming that addition is commutative). One obtains

$$(1 + 1)(a + b) = 1(a + b) + 1(a + b) = 1a + 1b + 1a + 1b = a + b + a + b$$

and

$$(1 + 1)(a + b) = (1 + 1)a + (1 + 1)b = 1a + 1a + 1b + 1b = a + a + b + b.$$

Since R is a group under addition, this implies $b + a = a + b$, i.e. that R under addition is necessarily commutative.

Fields are one of the most important examples of rings. Note that their definition below is just another formulation of the one given in Section 1.4.

Definition. A ring R with identity 1, where $1 \neq 0$, is called a *division ring* (or *skew field*) if every nonzero element $a \in R$ has a multiplicative inverse, i.e. there exists $b \in R$ such that $ab = ba = 1$. A commutative division ring is called a *field*.

More examples of rings follow.

Examples

(1) The simplest examples of rings are the *trivial rings* obtained by taking R to be any commutative group (denoting the group operation by $+$) and defining the multiplication $\times$ on R by: $a \times b = 0$ for all $a, b \in R$. It is easy to see that this multiplication defines a commutative ring. In particular, if $R = \{0\}$ is the trivial group, the resulting ring R is called the *zero ring*, denoted $R = 0$. Except for the zero ring, a trivial ring does not contain an identity ($R = 0$ is the only ring where $1 = 0$; we shall often exclude this ring by imposing the condition $1 \neq 0$). Although trivial rings have two binary operations, multiplication adds no new structure to the additive group and the theory of rings gives no information which could not already be obtained from (abelian) group theory.

(2) The ring of integers, $\mathbb{Z}$, under the usual operations of addition and multiplication is a commutative ring with identity (the integer 1). The ring axioms (as with the additive group axioms) follow from the basic axioms for the system of natural numbers. Note that under *multiplication* $\mathbb{Z} - \{0\}$ is *not* a group (in fact, there are very few multiplicative inverses to elements in this ring). We shall come back to the question of these inverses shortly.

(3) Similarly, the rational numbers, $\mathbb{Q}$, the real numbers, $\mathbb{R}$, and the complex numbers, $\mathbb{C}$, are commutative rings with identity (in fact they are fields). The ring axioms for each of these follow ultimately from the ring axioms for $\mathbb{Z}$. We shall verify this when we construct $\mathbb{Q}$ from $\mathbb{Z}$ (Section 7.5) and $\mathbb{C}$ from $\mathbb{R}$ (Example 1, Section 13.1); both of these constructions will be special cases of more general processes. The construction of $\mathbb{R}$ from $\mathbb{Q}$ (and subsequent verification of the ring axioms) is carried out in basic analysis texts.

(4) The quotient group $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring with identity (the element $\bar{1}$) under the operations of addition and multiplication of residue classes (frequently referred to as "modular arithmetic"). We saw that the additive abelian group axioms followed from the general principles of the theory of quotient groups (indeed this was the prototypical quotient group). We shall shortly prove that the remaining ring axioms (in particular, the fact that multiplication of residue classes is well defined) follow analogously from the general theory of quotient rings.

In all of the examples so far the rings have been commutative. Historically, one of the first noncommutative rings was discovered in 1843 by Sir William Rowell Hamilton (1805–1865). This ring, which is a division ring, was extremely influential in the subsequent development of mathematics and it continues to play an important role in certain areas of mathematics and physics.

(5) (The *(real) Hamilton Quaternions*) Let $\mathbb{H}$ be the collection of elements of the form $a + bi + cj + dk$ where $a, b, c, d \in \mathbb{R}$ are real numbers (loosely, "polynomials in $1, i, j, k$ with real coefficients") where addition is defined "componentwise" by

$$(a + bi + cj + dk) + (a' + b'i + c'j + d'k) = (a + a') + (b + b')i + (c + c')j + (d + d')k$$

and multiplication is defined by expanding $(a + bi + cj + dk)(a' + b'i + c'j + d'k)$ using the distributive law (being careful about the order of terms) and simplifying using the relations

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

(where the real number coefficients commute with i, j and k). For example,

$$\begin{aligned}(1 + i + 2j)(j + k) &= 1(j + k) + i(j + k) + 2j(j + k) = j + k + ij + ik + 2j^2 + 2j \\ &= j + k + k + (-j) + 2(-1) + 2(i) = -2 + 2i + 2k.\end{aligned}$$

The fact that $\mathbb{H}$ is a ring may be proved by a straightforward, albeit lengthy, check of the axioms (associativity of multiplication is particularly tedious). The Hamilton quaternions are a noncommutative ring with identity ($1 = 1 + 0i + 0j + 0k$). Similarly, one can define the ring of *rational* Hamilton Quaternions by taking a, b, c, d to be rational numbers above. Both the real and rational Hamilton Quaternions are *division rings*, where inverses of nonzero elements are given by

$$(a + bi + cj + dk)^{-1} = \frac{a - bi - cj - dk}{a^2 + b^2 + c^2 + d^2}.$$

(6) One important class of rings is obtained by considering rings of functions. Let X be any nonempty set and let A be any ring. The collection, R, of all (set) functions $f : X \to A$ is a ring under pointwise addition and multiplication of functions: $(f + g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$. Each ring axiom for R follows directly from the corresponding axiom for A. The ring R is commutative if and only if A is commutative and R has a 1 if and only if A has a 1 (in which case the 1 of R is necessarily the constant function 1 on X).

If X and A have more structure, we may form other rings of functions which respect those structures. For instance, if A is the ring of real numbers $\mathbb{R}$ and X is the closed interval $[0, 1]$ in $\mathbb{R}$ we may form the ring of all *continuous* functions from $[0, 1]$ to $\mathbb{R}$ (here we need basic limit theorems to guarantee that sums and products of continuous functions are continuous) — this is a commutative ring with 1.

(7) An example of a ring which does not have an identity is the ring $2\mathbb{Z}$ of even integers under usual addition and multiplication of integers (the sum and product of even integers is an even integer).

Another example which arises naturally in analysis is constructed as follows. A function $f : \mathbb{R} \to \mathbb{R}$ is said to have *compact support* if there are real numbers a, b (depending on f) such that $f(x) = 0$ for all $x \notin [a, b]$ (i.e. f is zero outside some bounded interval). The set of all functions $f : \mathbb{R} \to \mathbb{R}$ with compact support is a commutative ring without identity (since an identity could not have compact support). Similarly, the set of all continuous functions $f : \mathbb{R} \to \mathbb{R}$ with compact support is a commutative ring without identity.

In the next section we give three important ways of constructing "larger" rings from a given ring (analogous to Example 6 above) and thus greatly expand our list of examples. Before doing so we mention some basic properties of arbitrary rings. The ring $\mathbb{Z}$ is a good example to keep in mind, although this ring has a good deal more algebraic structure than a general ring (for example, it is commutative and has an identity). Nonetheless, its basic arithmetic holds for general rings as the following result shows.

Proposition 1. Let R be a ring. Then

(1) $0a = a0 = 0$ for all $a \in R$.

(2) $(-a)b = a(-b) = -(ab)$ for all $a, b \in R$ (recall $-a$ is the additive inverse of a).

(3) $(-a)(-b) = ab$ for all $a, b \in R$.

(4) if R has an identity 1, then the identity is unique and $-a = (-1)a$.

Proof: These all follow from the distributive laws and cancellation in the additive group R. For example, (1) follows from $0a = (0 + 0)a = 0a + 0a$. The equality $(-a)b = -(ab)$ in (2) follows from $ab + (-a)b = (a + (-a))b = 0b = 0$. The rest follow similarly and are left to the reader.

This proposition shows that because of the distributive laws the additive and multiplicative structures of a ring behave well with respect to one another, just as in the familiar example of the integers.

Unlike the integers, however, general rings may possess many elements that have multiplicative inverses or may have nonzero elements a and b whose product is zero. These two properties of elements, which relate to the multiplicative structure of a ring, are given special names.

Definition. Let R be a ring.

(1) A nonzero element a of R is called a *zero divisor* if there is a nonzero element b in R such that either $ab = 0$ or $ba = 0$.

(2) Assume R has an identity $1 \neq 0$. An element u of R is called a *unit* in R if there is some v in R such that $uv = vu = 1$. The set of units in R is denoted $R^\times$.

It is easy to see that the units in a ring R form a group under multiplication so $R^\times$ will be referred to as the *group of units* of R. In this terminology a *field is a commutative ring F with identity $1 \neq 0$ in which every nonzero element is a unit,* i.e. $F^\times = F - \{0\}$.

Observe that *a zero divisor can never be a unit*. Suppose for example that a is a unit in R and that $ab = 0$ for some nonzero b in R. Then $va = 1$ for some $v \in R$, so $b = 1b = (va)b = v(ab) = v0 = 0$, a contradiction. Similarly, if $ba = 0$ for some nonzero b then a cannot be a unit.

This shows in particular that fields contain no zero divisors.

Examples

(1) The ring $\mathbb{Z}$ of integers has no zero divisors and its only units are ± 1, i.e. $\mathbb{Z}^\times = \{\pm 1\}$. Note that every nonzero integer has an inverse in the *larger ring* $\mathbb{Q}$, so the property of being a unit depends on the ring in which an element is viewed.

(2) Let n be an integer ≥ 2. In the ring $\mathbb{Z}/n\mathbb{Z}$ the elements $\bar{u}$ for which u and n are relatively prime are units (we shall prove this in the next chapter). Thus our use of the notation $(\mathbb{Z}/n\mathbb{Z})^\times$ is consistent with the definition of the group of units in an arbitrary ring.

If, on the other hand, a is a nonzero integer and a is not relatively prime to n then we show that $\bar{a}$ is a zero divisor in $\mathbb{Z}/n\mathbb{Z}$. To see this let d be the g.c.d. of a and n and let $b = \dfrac{n}{d}$. By assumption $d > 1$ so $0 < b < n$, i.e. $\bar{b} \neq \bar{0}$. But by construction n divides ab, that is, $\overline{ab} = \bar{0}$ in $\mathbb{Z}/n\mathbb{Z}$. This shows that *every nonzero element of $\mathbb{Z}/n\mathbb{Z}$*

is either a unit or a zero divisor. Furthermore, every nonzero element is a unit if and only if every integer a in the range $0 < a < n$ is relatively prime to n. This happens if and only if n is a prime, i.e. *$\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is a prime*.

(3) If R is the ring of all functions from the closed interval [0,1] to $\mathbb{R}$ then the units of R are the functions that are not zero at any point (for such f its inverse is the function $\frac{1}{f}$). If f is not a unit and not zero then f is a zero divisor because if we define

$$g(x) = \begin{cases} 0, & \text{if } f(x) \neq 0 \\ 1, & \text{if } f(x) = 0 \end{cases}$$

then g is not the zero function but $f(x)g(x) = 0$ for all x.

(4) If R is the ring of all *continuous* functions from the closed interval [0,1] to $\mathbb{R}$ then the units of R are still the functions which are not zero at any point, but now there are functions that are neither units nor zero divisors. For instance, $f(x) = x - \frac{1}{2}$ has only one zero (at $x = \frac{1}{2}$) so f is not a unit. On the other hand, if $gf = 0$ then g must be zero for all $x \neq \frac{1}{2}$, and the only *continuous* function with this property is the zero function. Hence f is neither a unit nor a zero divisor. Similarly, no function with only a finite (or countable) number of zeros on [0,1] is a zero divisor. This ring also contains many zero divisors. For instance let

$$f(x) = \begin{cases} 0, & 0 \leq x \leq \frac{1}{2} \\ x - \frac{1}{2}, & \frac{1}{2} \leq x \leq 1 \end{cases}$$

and let $g(x) = f(1 - x)$. Then f and g are nonzero continuous functions whose product is the zero function.

Rings having some of the same characteristics as the integers $\mathbb{Z}$ are given a name:

Definition. A commutative ring with identity $1 \neq 0$ is called an *integral domain* if it has no zero divisors.

The absence of zero divisors in integral domains gives these rings a cancellation property:

Proposition 2. Assume a, b and c are elements of any ring with a not a zero divisor. If $ab = ac$, then either $a = 0$ or $b = c$ (i.e. if $a \neq 0$ we can cancel the a's). In particular, if a, b, c are any elements in an integral domain and $ab = ac$, then either $a = 0$ or $b = c$.

Proof: If $ab = ac$ then $a(b - c) = 0$ so either $a = 0$ or $b - c = 0$. The second statement follows from the first and the definition of an integral domain.

Corollary 3. Any finite integral domain is a field.

Proof: Let R be a finite integral domain and let a be a nonzero element of R. By the cancellation law the map $x \mapsto ax$ is an injective function. Since R is finite this map is also surjective. In particular, there is some $b \in R$ such that $ab = 1$, i.e. a is a unit in R. Since a was an arbitrary nonzero element, R is a field.

A remarkable result of Wedderburn is that a finite division ring is necessarily commutative, i.e. is a field. A proof of this theorem is outlined in the exercises at the end of Section 13.6.

In Section 5 below we study the relation between zero divisors and units in greater detail. We shall see that every nonzero element of a commutative ring that is not a zero divisor has a multiplicative inverse in some larger ring. This gives another perspective on the cancellation law in Proposition 2.

Having defined the notion of a ring, there is a natural notion of a subring.

Definition. A *subring* of the ring R is a subgroup of R that is closed under multiplication.

In other words, a subset S of a ring R is a subring if the operations of addition and multiplication in R when restricted to S give S the structure of a ring. To show that a subset of a ring R is a subring it suffices to check that it is *nonempty* and *closed under subtraction and under multiplication*.

Examples

A number of the examples above were also subrings.

(1) $\mathbb{Z}$ is a subring of $\mathbb{Q}$ and $\mathbb{Q}$ is a subring of $\mathbb{R}$. The property "is a subring of" is clearly transitive.

(2) $2\mathbb{Z}$ is a subring of $\mathbb{Z}$, as is $n\mathbb{Z}$ for any integer n. The ring $\mathbb{Z}/n\mathbb{Z}$ is not a subring (or a subgroup) of $\mathbb{Z}$ for any $n \geq 2$.

(3) The ring of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ is a subring of the ring of all functions from $\mathbb{R}$ to $\mathbb{R}$. The ring of all differentiable functions from $\mathbb{R}$ to $\mathbb{R}$ is a subring of both of these.

(4) $S = \mathbb{Z} + \mathbb{Z}i + \mathbb{Z}j + \mathbb{Z}k$, the *integral Quaternions*, form a subring of either the real or the rational Quaternions — it is easy to check that multiplying two such quaternions together gives another quaternion with integer coefficients. This ring (which is not a division ring) can be used to give proofs for a number of results in number theory.

(5) If R is a subring of a field F that contains the identity of F then R is an integral domain. The converse of this is also true, namely any integral domain is contained in a field (cf. Section 5).

EXERCISES

Let R be a ring with 1.

1. Show that $(-1)^2 = 1$ in R.

2. Prove that if u is a unit in R then so is $-u$.

3. Let R be a ring with identity and let S be a subring of R containing the identity. Prove that if u is a unit in S then u is a unit in R. Show by example that the converse is false.

4. Prove that the intersection of any nonempty collection of subrings of a ring is also a subring.

5. Decide which of the following are subrings of $\mathbb{Q}$:

(a) the set of all rational numbers with odd denominators (when written in lowest terms)

(b) the set of all rational numbers with even denominators (when written in lowest terms)

(c) the set of nonnegative rational numbers

(d) the set of squares of rational numbers

(e) the set of all rational numbers with odd numerators (when written in lowest terms)
(f) the set of all rational numbers with even numerators (when written in lowest terms).

6. Decide which of the following are subrings of the ring of all functions from the closed interval [0,1] to $\mathbb{R}$:
   (a) the set of all functions $f(x)$ such that $f(q) = 0$ for all $q \in \mathbb{Q} \cap [0, 1]$
   (b) the set of all polynomial functions
   (c) the set of all functions which have only a finite number of zeros, together with the zero function
   (d) the set of all functions which have an infinite number of zeros
   (e) the set of all functions f such that $\lim_{x \to 1^+} f(x) = 0$
   (f) the set of all rational linear combinations of the functions $\sin nx$ and $\cos mx$, where $m, n \in \{0, 1, 2, \ldots\}$.
7. The *center* of a ring R is $\{z \in R \mid zr = rz \text{ for all } r \in R\}$ (i.e. is the set of all elements which commute with every element of R). Prove that the center of a ring is a subring that contains the identity. Prove that the center of a division ring is a field.
8. Describe the center of the real Hamilton quaternions $\mathbb{H}$. Prove that $\{a + bi \mid a, b \in \mathbb{R}\}$ is a subring of $\mathbb{H}$ which is a field but is not contained in the center of $\mathbb{H}$.
9. For a fixed element $a \in R$ define $C(a) = \{r \in R \mid ra = ar\}$. Prove that $C(a)$ is a subring of R containing a. Prove that the center of R is the intersection of the subrings $C(a)$ over all $a \in R$.
10. Prove that if D is a division ring then $C(a)$ is a division ring for all $a \in D$ (cf. the preceding exercise).
11. Prove that if R is an integral domain and $x^2 = 1$ for some $x \in R$ then $x = \pm 1$.
12. Prove that any subring of a field which contains the identity is an integral domain.
13. An element x in R is called *nilpotent* if $x^m = 0$ for some $m \in \mathbb{Z}^+$.
    (a) Show that if $n = a^k b$ for some integers a and b then $\overline{ab}$ is a nilpotent element of $\mathbb{Z}/n\mathbb{Z}$.
    (b) If $a \in \mathbb{Z}$ is an integer, show that the element $\bar{a} \in \mathbb{Z}/n\mathbb{Z}$ is nilpotent if and only if every prime divisor of n is also a divisor of a. In particular, determine the nilpotent elements of $\mathbb{Z}/72\mathbb{Z}$ explicitly.
    (c) Let R be the ring of functions from a nonempty set X to a field F. Prove that R contains no nonzero nilpotent elements.
14. Let x be a nilpotent element of the commutative ring R (cf. the preceding exercise).
    (a) Prove that x is either zero or a zero divisor.
    (b) Prove that rx is nilpotent for all $r \in R$.
    (c) Prove that $1 + x$ is a unit in R.
    (d) Deduce that the sum of a nilpotent element and a unit is a unit.
15. A ring R is called a *Boolean ring* if $a^2 = a$ for all $a \in R$. Prove that every Boolean ring is commutative.
16. Prove that the only Boolean ring that is an integral domain is $\mathbb{Z}/2\mathbb{Z}$.
17. Let R and S be rings. Prove that the direct product $R \times S$ is a ring under componentwise addition and multiplication. Prove that $R \times S$ is commutative if and only if both R and S are commutative. Prove that $R \times S$ has an identity if and only if both R and S have identities.
18. Prove that $\{(r, r) \mid r \in R\}$ is a subring of $R \times R$.
19. Let I be any nonempty index set and let R_i be a ring for each $i \in I$. Prove that the direct product $\prod_{i \in I} R_i$ is a ring under componentwise addition and multiplication.
20. Let R be the collection of sequences $(a_1, a_2, a_3, \ldots)$ of integers $a_1, a_2, a_3, \ldots$ where all but finitely many of the a_i are 0 (called the *direct sum* of infinitely many copies of $\mathbb{Z}$). Prove

that R is a ring under componentwise addition and multiplication which does not have an identity.

21. Let X be any nonempty set and let $\mathcal{P}(X)$ be the set of all subsets of X (the *power set* of X). Define addition and multiplication on $\mathcal{P}(X)$ by

$$A + B = (A - B) \cup (B - A) \qquad \text{and} \qquad A \times B = A \cap B$$

i.e. addition is symmetric difference and multiplication is interesction.

(a) Prove that $\mathcal{P}(X)$ is a ring under these operations ($\mathcal{P}(X)$ and its subrings are often referred to as *rings of sets*).

(b) Prove that this ring is commutative, has an identity and is a Boolean ring.

22. Give an example of an infinite Boolean ring.

23. Let D be an integer that is not a perfect square in $\mathbb{Z}$ and define the following subset of $\mathbb{C}$:

$$\mathbb{Z}[\sqrt{D}] = \{a + b\sqrt{D} \mid a, b \in \mathbb{Z}\}.$$

(a) Prove that $\mathbb{Z}[\sqrt{D}]$ is a subring of $\mathbb{C}$ (so if $D > 0$ then $\mathbb{Z}[\sqrt{D}]$ is a subring of $\mathbb{R}$).

(b) Define $N : \mathbb{Z}[\sqrt{D}] \to \mathbb{Z}$ by $N(a + b\sqrt{D}) = a^2 - Db^2$ (the map N is called a *norm*). Prove that $N(xy) = N(x)N(y)$ for all $x, y \in \mathbb{Z}[\sqrt{D}]$. Deduce that if u is a unit in $\mathbb{Z}[\sqrt{D}]$ then $N(u) = \pm 1$.

(c) Prove that the inverse in $\mathbb{C}$ of the element $a + b\sqrt{D}$ is $\dfrac{a - b\sqrt{D}}{a^2 - Db^2}$. Deduce that if $u \in \mathbb{Z}[\sqrt{D}]$ and $N(u) = \pm 1$, then u is a unit.

(d) Prove that if $D < -1$ then $\mathbb{Z}[\sqrt{D}]^\times = \{\pm 1\}$.

(e) Prove that the group of units in $\mathbb{Z}[i]$ is $\{\pm 1, \pm i\}$, where $i = \sqrt{-1}$.

(f) Prove that the element $1 + \sqrt{2}$ is a unit in $\mathbb{Z}[\sqrt{2}]$. Show that $\mathbb{Z}[\sqrt{2}]^\times$ is an infinite group. [Show that $1 + \sqrt{2}$ has infinite order.]

24. Let D be an integer that is not a perfect square in $\mathbb{Z}$ and define the following subset of $\mathbb{C}$:

$$\mathbb{Q}[\sqrt{D}] = \{a + b\sqrt{D} \mid a, b \in \mathbb{Q}\}.$$

Prove that $\mathbb{Q}[\sqrt{D}]$ is a field.

25. Let I be the ring of integral Hamilton quaternions and define

$$N : I \to \mathbb{Z} \qquad \text{by} \qquad N(a + bi + cj + dk) = a^2 + b^2 + c^2 + d^2$$

(the map N is called a *norm*).

(a) Prove that $N(\alpha) = \alpha\overline{\alpha}$ for all $\alpha \in I$, where if $\alpha = a + bi + cj + dk$ then $\overline{\alpha} = a - bi - cj - dk$.

(b) Prove that $N(\alpha\beta) = N(\alpha)N(\beta)$ for all $\alpha, \beta \in I$.

(c) Prove that an element of I is a unit if and only if it has norm ± 1. Show that $I^\times$ is isomorphic to the quaternion group of order 8. [The inverse in the ring of rational quaternions of a nonzero element α is $\dfrac{\overline{\alpha}}{N(\alpha)}$.]

26. Let K be a field. A *discrete valuation* on K is a function $\nu : K^\times \to \mathbb{Z}$ satisfying

(i) $\nu(ab) = \nu(a) + \nu(b)$ (i.e. ν is a homomorphism from the multiplicative group of nonzero elements of K to $\mathbb{Z}$),

(ii) ν is surjective, and

(iii) $\nu(x + y) \geq \min\{\nu(x), \nu(y)\}$ for all $x, y \in K^\times$ with $x + y \neq 0$.

The set $R = \{x \in K^\times \mid \nu(x) \geq 0\} \cup \{0\}$ is called the *valuation ring* of ν.

(a) Prove that R is a subring of K which contains the identity. (In general, a ring R is called

a *discrete valuation ring* if there is some field K and some discrete valuation ν on K such that R is the valuation ring of ν).

(b) Prove that for each nonzero element $x \in K$ either x or x^{-1} is in R.

(c) Prove that an element x is a unit of R if and only if $\nu(x) = 0$.

27. A specific example of a discrete valuation ring (cf. the preceding exercise) is obtained when p is a prime, $K = \mathbb{Q}$ and

$$\nu_p : \mathbb{Q}^\times \to \mathbb{Z} \qquad \text{by} \qquad \nu_p(\frac{a}{b}) = \alpha \quad \text{where } \frac{a}{b} = p^\alpha \frac{c}{d}, \quad p \nmid c \text{ and } p \nmid d.$$

Prove that the corresponding valuation ring R is the ring of all rational numbers whose denominators are relatively prime to p. Describe the units of this valuation ring.

28. Let R be a ring with $1 \neq 0$. A nonzero element a is called a *left zero divisor* in R if there is a nonzero element $x \in R$ such that $ax = 0$. Symmetrically, $b \neq 0$ is a *right zero divisor* if there is a nonzero $y \in R$ such that $yb = 0$ (so a zero divisor is an element which is either a left or a right zero divisor). An element $u \in R$ has a *left inverse* in R if there is some $s \in R$ such that $su = 1$. Symmetrically, v has a *right inverse* if $vt = 1$ for some $t \in R$.

(a) Prove that u is a unit if and only if it has both a right and a left inverse (i.e. u must have a two-sided inverse).

(b) Prove that if u has a right inverse then u is not a right zero divisor.

(c) Prove that if u has more than one right inverse then u is a left zero divisor.

(d) Prove that if R is a finite ring then every element that has a right inverse is a unit (i.e. has a two-sided inverse).

29. Let A be any commutative ring with identity $1 \neq 0$. Let R be the set of all group homomorphisms of the additive group A to itself with addition defined as pointwise addition of functions and multiplication defined as function composition. Prove that these operations make R into a ring with identity. Prove that the units of R are the group automorphisms of A (cf. Exercise 20, Section 1.6).

30. Let $A = \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z} \times \cdots$ be the direct product of copies of $\mathbb{Z}$ indexed by the positive integers (so A is a ring under componentwise addition and multiplication) and let R be the ring of all group homomorphisms from A to itself as described in the preceding exercise. Let φ be the element of R defined by $\varphi(a_1, a_2, a_3, \dots) = (a_2, a_3, \dots)$. Let ψ be the element of R defined by $\psi(a_1, a_2, a_3, \dots) = (0, a_1, a_2, a_3, \dots)$.

(a) Prove that $\varphi\psi$ is the identity of R but $\psi\varphi$ is not the identity of R (i.e. ψ is a *right inverse* for φ but not a left inverse.

(b) Exhibit infinitely many right inverses for φ.

(c) Find a nonzero element π in R such that $\varphi\pi = 0$ but $\pi\varphi \neq 0$.

(d) Prove that there is no nonzero element $\lambda \in R$ such that $\lambda\varphi = 0$ (i.e. φ is a left zero divisor but not a right zero divisor).

7.2 EXAMPLES: POLYNOMIAL RINGS, MATRIX RINGS, AND GROUP RINGS

We introduce here three important types of rings: polynomial rings, matrix rings, and group rings. We shall see in the course of the text that these three classes of rings are often related. For example, we shall see in Part V that the group ring of a group G over the complex numbers $\mathbb{C}$ is a direct product of matrix rings over $\mathbb{C}$.

These rings also have many important applications, in addition to being interesting in their own right. In Part III we shall use polynomial rings to prove some

classification theorems for matrices which, in particular, determine when a matrix is similar to a diagonal matrix. In Part V we shall use group rings to study group actions and to prove some additional important classification theorems.

Polynomial Rings

Fix a commutative ring R with identity. We define the ring of polynomials in a form which may already be familiar, at least for polynomials with real coefficients. A definition in terms of Cartesian products is given in the Appendix. Let x be an indeterminate. The formal sum

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with $n \geq 0$ and each $a_i \in R$ is called a *polynomial* in x with coefficients a_i in R. If $a_n \neq 0$, then the polynomial is said to be of *degree* n and $a_n x^n$ is called the *leading term*. The polynomial is *monic* if $a_n = 1$. The set of all such polynomials is called the ring of *polynomials in the variable x with coefficients in R* and will be denoted $R[x]$.

The operations of addition and multiplication which make $R[x]$ into a ring are the same operations familiar from elementary algebra: addition is "componentwise"

$$\begin{aligned}(a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0) + (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0)\\ = (a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_1 + b_1)x + (a_0 + b_0)\end{aligned}$$

(here a_n or b_n may be zero in order for addition of polynomials of different degrees to be defined). Multiplication is performed by first defining $(ax^i)(bx^j) = abx^{i+j}$ for *monomials* (polynomials with only one nonzero term) and then extending to all polynomials by the distributive laws (usually referred to as "expanding out and collecting like terms"):

$$\begin{aligned}&(a_0 + a_1 x + a_2 x^2 + \ldots) \times (b_0 + b_1 x + b_2 x^2 + \ldots)\\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \ldots\end{aligned}$$

(in general, the coefficient of x^k in the product will be $\sum_{i=0}^{k} a_i b_{k-i}$). These operations make sense since R is a ring so the sums and products of the coefficients are defined. An easy verification proves that $R[x]$ is indeed a ring with these definitions of addition and multiplication.

The ring R appears in $R[x]$ as the *constant polynomials*. Note that by definition of the multiplication, $R[x]$ is a *commutative ring with identity* (the identity 1 from R).

The coefficient ring R above was assumed to be a commutative ring since that is the situation we shall be primarily interested in, but note that the definition of the addition and multiplication in $R[x]$ above would be valid even if R were not commutative or did not have an identity. If the coefficient ring R is the integers $\mathbb{Z}$ (respectively, the rationals $\mathbb{Q}$) the polynomial ring $\mathbb{Z}[x]$ (respectively, $\mathbb{Q}[x]$) is the ring of polynomials with integer (rational) coefficients familiar from elementary algebra.

Another example is the polynomial ring $\mathbb{Z}/3\mathbb{Z}[x]$ of polynomials in x with coefficients in $\mathbb{Z}/3\mathbb{Z}$. This ring consists of nonnegative powers of x with coefficients 0, 1, and 2 with calculations on the coefficients performed modulo 3. For example, if

$$P(x) = x^2 + 2x + 1 \qquad \text{and} \qquad Q(x) = x^3 + x + 2$$

then

$$P(x) + Q(x) = x^3 + x^2$$

and

$$P(x)Q(x) = x^5 + 2x^4 + 2x^3 + x^2 + 2x + 2.$$

The ring in which the coefficients are taken makes a substantial difference in the behavior of polynomials. For example, the polynomial $x^2 + 1$ is not a perfect square in the polynomial ring $\mathbb{Z}[x]$, but *is* a perfect square in the polynomial ring $\mathbb{Z}/2\mathbb{Z}[x]$, since $(x + 1)^2 = x^2 + 2x + 1 = x^2 + 1$ in this ring.

Proposition 4. Let R be an integral domain and let $p(x), q(x) \in R[x]$. Then

(1) degree $p(x)q(x)$ = degree $p(x)$ + degree $q(x)$,

(2) the units of $R[x]$ are just the units of R,

(3) $R[x]$ is an integral domain.

Proof: If R has no zero divisors then neither does $R[x]$; if $p(x)$ and $q(x)$ are polynomials with leading terms $a_n x^n$ and $b_m x^m$, respectively, then the leading term of $p(x)q(x)$ is $a_n b_m x^{n+m}$, and $a_n b_m \neq 0$. This proves (3) and also verifies (1). If $p(x)$ is a unit, say $p(x)q(x) = 1$ in $R[x]$, then degree $p(x)$ + degree $q(x) = 0$, so both $p(x)$ and $q(x)$ are elements of R, hence are units in R since their product is 1. This proves (2).

If the ring R has zero divisors then so does $R[x]$, because $R \subset R[x]$. Also, if $f(x)$ is a zero divisor in $R[x]$ (i.e. $f(x)g(x) = 0$ for some nonzero $g(x) \in R[x]$) then in fact $cf(x) = 0$ for some nonzero $c \in R$ (cf. Exercise 2).

If S is a subring of R then $S[x]$ is a subring of $R[x]$. For instance, $\mathbb{Z}[x]$ is a subring of $\mathbb{Q}[x]$. Some other examples of subrings of $R[x]$ are the set of all polynomials in x^2 (i.e. in which only even powers of x appear) and the set of all polynomials with zero constant term (the latter subring does not have an identity).

Polynomial rings, particularly those over fields, will be studied extensively in Chapter 9.

Matrix Rings

Fix an arbitrary ring R and let n be a positive integer. Let $M_n(R)$ be the set of all $n \times n$ *matrices with entries from* R. The element (a_{ij}) of $M_n(R)$ is an $n \times n$ square array of elements of R whose entry in row i and column j is $a_{ij} \in R$. The set of matrices becomes a ring under the usual rules by which matrices of real numbers are added and multiplied. Addition is componentwise: the i, j entry of the matrix $(a_{ij}) + (b_{ij})$ is $a_{ij} + b_{ij}$. The i, j entry of the matrix product $(a_{ij}) \times (b_{ij})$ is $\sum_{k=1}^{n} a_{ik} b_{kj}$ (note that these matrices need to be square in order that multiplication of any two elements be defined). It is a straightforward calculation to check that these operations make $M_n(R)$ into a ring. When R is a field we shall prove that $M_n(R)$ is a ring by less computational means in Part III.

Note that if R is any nontrivial ring (even a commutative one) and $n \geq 2$ then $M_n(R)$ is *not commutative*: if $ab \neq 0$ in R let A be the matrix with a in position 1,1 and

zeros elsewhere and let B be the matrix with b in position 1,2 and zeros elsewhere; then ab is the (nonzero) entry in position 1,2 of AB whereas BA is the zero matrix.

These two matrices also show that $M_n(R)$ has zero divisors for all nonzero rings R whenever $n \geq 2$.

An element (a_{ij}) of $M_n(R)$ is called a *scalar matrix* if for some $a \in R$, $a_{ii} = a$ for all $i \in \{1, \ldots, n\}$ and $a_{ij} = 0$ for all $i \neq j$ (i.e. all diagonal entries equal a and all off-diagonal entries are 0). The set of scalar matrices is a subring of $M_n(R)$. This subring is a copy of R (i.e. is "isomorphic" to R): if the matrix A has the element a along the main diagonal and the matrix B has the element b along the main diagonal then the matrix $A + B$ has $a + b$ along the diagonal and AB has ab along the diagonal (and all other entries 0). If R is commutative, the scalar matrices commute with all elements of $M_n(R)$. If R has a 1, then the scalar matrix with 1's down the diagonal (the $n \times n$ *identity matrix*) is the 1 of $M_n(R)$. In this case the units in $M_n(R)$ are the invertible $n \times n$ matrices and the group of units is denoted $GL_n(R)$, the *general linear group* of degree n over R.

If S is a subring of R then $M_n(S)$ is a subring of $M_n(R)$. For instance $M_n(\mathbb{Z})$ is a subring of $M_n(\mathbb{Q})$ and $M_n(2\mathbb{Z})$ is a subring of both of these. Another example of a subring of $M_n(R)$ is the set of *upper triangular* matrices: $\{(a_{ij}) \mid a_{pq} = 0 \text{ whenever } p > q\}$ (the set of matrices all of whose entries below the main diagonal are zero) — one easily checks that the sum and product of upper triangular matrices is upper triangular.

Group Rings

Fix a commutative ring R with identity $1 \neq 0$ and let $G = \{g_1, g_2, \ldots, g_n\}$ be any finite group with group operation written multiplicatively. Define the *group ring*, RG, of G with coefficients in R to be the set of all formal sums

$$a_1g_1 + a_2g_2 + \cdots + a_ng_n \qquad a_i \in R, \quad 1 \leq i \leq n.$$

If g_1 is the identity of G we shall write a_1g_1 simply as a_1. Similarly, we shall write the element $1g$ for $g \in G$ simply as g.

Addition is defined "componentwise"

$$\begin{aligned}(a_1g_1 + a_2g_2 + \cdots + a_ng_n) &+ (b_1g_1 + b_2g_2 + \cdots + b_ng_n)\\ &= (a_1 + b_1)g_1 + (a_2 + b_2)g_2 + \cdots + (a_n + b_n)g_n.\end{aligned}$$

Multiplication is performed by first defining $(ag_i)(bg_j) = (ab)g_k$, where the product ab is taken in R and $g_ig_j = g_k$ is the product in the group G. This product is then extended to all formal sums by the distributive laws so that the coefficient of g_k in the product $(a_1g_1 + \cdots + a_ng_n) \times (b_1g_1 + \cdots + b_ng_n)$ is $\sum_{g_ig_j = g_k} a_ib_j$. It is straightforward to check that these operations make RG into a ring (again, commutativity of R is not needed). The associativity of multiplication follows from the associativity of the group operation in G. The ring RG is commutative if and only if G is a commutative group.

Example

Let $G = D_8$ be the dihedral group of order 8 with the usual generators r, s ($r^4 = s^2 = 1$ and $rs = sr^{-1}$) and let $R = \mathbb{Z}$. The elements $\alpha = r + r^2 - 2s$ and $\beta = -3r^2 + rs$ are

typical members of $\mathbb{Z}D_8$. Their sum and product are then

$$\begin{aligned}
\alpha + \beta &= r - 2r^2 - 2s + rs \\
\alpha\beta &= (r + r^2 - 2s)(-3r^2 + rs) \\
&= r(-3r^2 + rs) + r^2(-3r^2 + rs) - 2s(-3r^2 + rs) \\
&= -3r^3 + r^2s - 3 + r^3s + 6r^2s - 2r^3 \\
&= -3 - 5r^3 + 7r^2s + r^3s.
\end{aligned}$$

The ring R appears in RG as the "constant" formal sums i.e. the R–multiples of the identity of G (note that the definition of the addition and multiplication in RG restricted to these elements is just the addition and multiplication in R). These elements of R commute with all elements of RG. The identity of R is the identity of RG.

The group G also appears in RG (the element g_i appears as $1g_i$ — for example, $r, s \in D_8$ are also elements of the group ring $\mathbb{Z}D_8$ above) — multiplication in the ring RG restricted to G is just the group operation. In particular, each element of G has a multiplicative inverse in the ring RG (namely, its inverse in G). This says that G is a *subgroup of the group of units of* RG.

If $|G| > 1$ then RG always has zero divisors. For example, let g be any element of G of order $m > 1$. Then

$$(1 - g)(1 + g + \cdots + g^{m-1}) = 1 - g^m = 1 - 1 = 0$$

so $1 - g$ is a zero divisor (note that by definition of RG neither of the formal sums in the above product is zero).

If S is a subring of R then SG is a subring of RG. For instance, $\mathbb{Z}G$ (called the *integral group ring* of G) is a subring of $\mathbb{Q}G$ (the *rational group ring* of G). Furthermore, if H is a subgroup of G then RH is a subring of RG. The set of all elements of RG whose coefficients sum to zero is a subring (without identity). If $|G| > 1$, the set of elements with zero "constant term" (i.e. the coefficient of the identity of G is zero) is *not* a subring (it is not closed under multiplication).

Note that the group ring $\mathbb{R}Q_8$ is *not* the same ring as the Hamilton Quaternions $\mathbb{H}$ even though the latter contains a copy of the quaternion group Q_8 (under multiplication). One difference is that the unique element of order 2 in Q_8 (usually denoted by -1) is not the additive inverse of 1 in $\mathbb{R}Q_8$. In other words, if we temporarily denote the identity of the group Q_8 by g_1 and the unique element of order 2 by g_2, then $g_1 + g_2$ is not zero in $\mathbb{R}Q_8$, whereas $1 + (-1)$ is zero in $\mathbb{H}$. Furthermore, as noted above, the group ring $\mathbb{R}Q_8$ contains zero divisors, hence is not a division ring.

Group rings over fields will be studied extensively in Chapter 15.

EXERCISES

Let R be a commutative ring with 1.

1. Let $p(x) = 2x^3 - 3x^2 + 4x - 5$ and let $q(x) = 7x^3 + 33x - 4$. In each of parts (a), (b) and (c) compute $p(x) + q(x)$ and $p(x)q(x)$ under the assumption that the coefficients of the two given polynomials are taken from the specified ring (where the integer coefficients are taken mod n in parts (b) and (c)):
   (a) $R = \mathbb{Z}$, (b) $R = \mathbb{Z}/2\mathbb{Z}$, (c) $R = \mathbb{Z}/3\mathbb{Z}$.

2. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be an element of the polynomial ring $R[x]$. Prove that $p(x)$ is a zero divisor in $R[x]$ if and only if there is a nonzero $b \in R$ such that $bp(x) = 0$. [Let $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_0$ be a nonzero polynomial of minimal degree such that $g(x)p(x) = 0$. Show that $b_m a_n = 0$ and so $a_n g(x)$ is a polynomial of degree less than m that also gives 0 when multiplied by $p(x)$. Conclude that $a_n g(x) = 0$. Apply a similar argument to show by induction on i that $a_{n-i}g(x) = 0$ for $i = 0, 1, \ldots, n$, and show that this implies $b_0 p(x) = 0$.]

3. Define the set $R[[x]]$ of *formal power series* in the indeterminate x with coefficients from R to be all formal infinite sums

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots.$$

Define addition and multiplication of power series in the same way as for power series with real or complex coefficients, i.e. extend polynomial addition and multiplication to power series as though they were "polynomials of infinite degree":

$$\sum_{n=0}^{\infty} a_n x^n \quad + \quad \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} (a_n + b_n) x^n$$

$$\sum_{n=0}^{\infty} a_n x^n \quad \times \quad \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} \Big(\sum_{k=0}^{n} a_k b_{n-k}\Big) x^n.$$

(The term "formal" is used here to indicate that convergence is not considered, so that formal power series need not represent functions on R.)

(a) Prove that $R[[x]]$ is a commutative ring with 1.

(b) Show that $1 - x$ is a unit in $R[[x]]$ with inverse $1 + x + x^2 + \cdots$.

(c) Prove that $\sum_{n=0}^{\infty} a_n x^n$ is a unit in $R[[x]]$ if and only if a_0 is a unit in R.

4. Prove that if R is an integral domain then the ring of formal power series $R[[x]]$ is also an integral domain.

5. Let F be a field and define the ring $F((x))$ of *formal Laurent series* with coefficients from F by

$$F((x)) = \{ \sum_{n \geq N}^{\infty} a_n x^n \mid a_n \in F \text{ and } N \in \mathbb{Z} \}$$

(so every element of $F((x))$ is a power series in positive powers of x plus a polynomial in $\frac{1}{x}$, i.e. each element of $F((x))$ has only a finite number of terms with negative powers of x).

(a) Prove that $F((x))$ is a field.

(b) Define the map

$$\nu : F((x))^{\times} \to \mathbb{Z} \qquad \text{by} \qquad \nu(\sum_{n \geq N}^{\infty} a_n x^n) = N$$

where a_N is the first nonzero coefficient of the series (i.e. N is the "order of zero or pole of the series at 0"). Prove that ν is a discrete valuation on $F((x))$ whose discrete valuation ring is $F[[x]]$, the ring of formal power series (cf. Exercise 26, Section 1).

6. Let S be a ring with identity $1 \neq 0$. Let $n \in \mathbb{Z}^+$ and let A be an $n \times n$ matrix with entries from S whose i, j entry is a_{ij}. Let E_{ij} be the element of $M_n(S)$ whose i, j entry is 1 and whose other entries are all 0.

(a) Prove that $E_{ij}A$ is the matrix whose i^{th} row equals the jth row of A and all other rows are zero.

(b) Prove that AE_{ij} is the matrix whose j^{th} column equals the ith column of A and all other columns are zero.

(c) Deduce that $E_{pq}AE_{rs}$ is the matrix whose p, s entry is a_{qr} and all other entries are zero.

7. Prove that the center of the ring $M_n(R)$ is the set of scalar matrices (cf. Exercise 7, Section 1). [Use the preceding exercise.]

8. Let S be any ring and let $n \geq 2$ be an integer. Prove that if A is any strictly upper triangular matrix in $M_n(S)$ then $A^n = 0$ (a strictly upper triangular matrix is one whose entries on and below the main diagonal are all zero).

9. Let $\alpha = r + r^2 - 2s$ and $\beta = -3r^2 + rs$ be the two elements of the integral group ring $\mathbb{Z}D_8$ described in this section. Compute the following elements of $\mathbb{Z}D_8$:
(a) $\beta\alpha$, (b) α^2, (c) $\alpha\beta - \beta\alpha$, (d) $\beta\alpha\beta$.

10. Consider the following elements of the integral group ring $\mathbb{Z}S_3$:

$$\alpha = 3(1\ 2) - 5(2\ 3) + 14(1\ 2\ 3) \qquad \text{and} \qquad \beta = 6(1) + 2(2\ 3) - 7(1\ 3\ 2)$$

(where (1) is the identity of S_3). Compute the following elements:
(a) $\alpha + \beta$, (b) $2\alpha - 3\beta$, (c) $\alpha\beta$, (d) $\beta\alpha$, (e) α^2.

11. Repeat the preceding exercise under the assumption that the coefficients of α and β are in $\mathbb{Z}/3\mathbb{Z}$ (i.e. $\alpha, \beta \in \mathbb{Z}/3\mathbb{Z}S_3$).

12. Let $G = \{g_1, \ldots, g_n\}$ be a finite group. Prove that the element $N = g_1 + g_2 + \ldots + g_n$ is in the center of the group ring RG (cf. Exercise 7, Section 1).

13. Let $\mathcal{K} = \{k_1, \ldots, k_m\}$ be a conjugacy class in the finite group G.

(a) Prove that the element $K = k_1 + \ldots + k_m$ is in the center of the group ring RG (cf. Exercise 7, Section 1). [Check that $g^{-1}Kg = K$ for all $g \in G$.]

(b) Let $\mathcal{K}_1, \ldots, \mathcal{K}_r$ be the conjugacy classes of G and for each $\mathcal{K}_i$ let K_i be the element of RG that is the sum of the members of $\mathcal{K}_i$. Prove that an element $\alpha \in RG$ is in the center of RG if and only if $\alpha = a_1K_1 + a_2K_2 + \cdots + a_rK_r$ for some $a_1, a_2, \ldots, a_r \in R$.

7.3 RING HOMOMORPHISMS AND QUOTIENT RINGS

A ring homomorphism is a map from one ring to another that respects the additive and multiplicative structures:

Definition. Let R and S be rings.

(1) A *ring homomorphism* is a map $\varphi : R \to S$ satisfying

(i) $\varphi(a + b) = \varphi(a) + \varphi(b)$ for all $a, b \in R$ (so φ is a group homomorphism on the additive groups) and

(ii) $\varphi(ab) = \varphi(a)\varphi(b)$ for all $a, b \in R$.

(2) The *kernel* of the ring homomorphism φ, denoted $\ker \varphi$, is the set of elements of R that map to 0 in S (i.e. the kernel of φ viewed as a homomorphism of additive groups).

(3) A bijective ring homomorphism is called an *isomorphism*.

If the context is clear we shall simply use the term "homomorphism" instead of "ring homomorphism." Similarly, if A and B are rings, $A \cong B$ will always mean an isomorphism of rings unless otherwise stated.

Examples

(1) The map $\varphi : \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ defined by sending an even integer to 0 and an odd integer to 1 is a ring homomorphism. The map is additive since the sum of two even or odd integers is even and the sum of an even integer and an odd integer is odd. The map is multiplicative since the product of two odd integers is odd and the product of an even integer with any integer is even. The kernel of φ (the fiber of φ above $0 \in \mathbb{Z}/2\mathbb{Z}$) is the set of even integers. The fiber of φ above $1 \in \mathbb{Z}/2\mathbb{Z}$ is the set of odd integers.

(2) For $n \in \mathbb{Z}$ the maps $\varphi_n : \mathbb{Z} \to \mathbb{Z}$ defined by $\varphi_n(x) = nx$ are *not* in general ring homomorphisms because $\varphi_n(xy) = nxy$ whereas $\varphi_n(x)\varphi_n(y) = nxny = n^2xy$. Hence φ_n is a ring homomorphism only when $n^2 = n$, i.e. $n = 0, 1$. Note however that φ_n is always a *group homomorphism* on the additive groups. Thus care should be exercised when dealing with rings to be sure to check that *both* ring operations are preserved. Note that φ_0 is the zero homomorphism and φ_1 is the identity homomorphism.

(3) Let $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ be the map from the ring of polynomials in x with rational coefficients to the rationals defined by $\varphi(p(x)) = p(0)$ (i.e. mapping the polynomial to its constant term). Then φ is a ring homomorphism since the constant term of the sum of two polynomials is the sum of their constant terms and the constant term of the product of two polynomials is the product of their constant terms. The fiber above $a \in \mathbb{Q}$ consists of the set of polynomials with a as constant term. In particular, the kernel of φ consists of the polynomials with constant term 0.

Proposition 5. Let R and S be rings and let $\varphi : R \to S$ be a homomorphism.

(1) The image of φ is a subring of S.

(2) The kernel of φ is a subring of R. Furthermore, if $\alpha \in \ker \varphi$ then $r\alpha$ and $\alpha r \in \ker \varphi$ for every $r \in R$, i.e. $\ker \varphi$ is closed under multiplication by elements from R.

Proof: (1) If $s_1, s_2 \in \operatorname{im} \varphi$ then $s_1 = \varphi(r_1)$ and $s_2 = \varphi(r_2)$ for some $r_1, r_2 \in R$. Then $\varphi(r_1 - r_2) = s_1 - s_2$ and $\varphi(r_1 r_2) = s_1 s_2$. This shows $s_1 - s_2, s_1 s_2 \in \operatorname{im} \varphi$, so the image of φ is closed under subtraction and under multiplication, hence is a subring of S.

(2) If $\alpha, \beta \in \ker \varphi$ then $\varphi(\alpha) = \varphi(\beta) = 0$. Hence $\varphi(\alpha - \beta) = 0$ and $\varphi(\alpha\beta) = 0$, so $\ker \varphi$ is closed under subtraction and under multiplication, so is a subring of R. Similarly, for any $r \in R$ we have $\varphi(r\alpha) = \varphi(r)\varphi(\alpha) = \varphi(r)0 = 0$, and $\varphi(\alpha r) = \varphi(\alpha)\varphi(r) = 0\varphi(r) = 0$, so $r\alpha, \alpha r \in \ker \varphi$.

In the case of a homomorphism φ of groups we saw that the fibers of the homomorphism have the structure of a group naturally isomorphic to the image of φ, which led to the notion of a quotient group by a normal subgroup. An analogous result is true for a homomorphism of rings.

Let $\varphi : R \to S$ be a ring homomorphism with kernel I. Since R and S are in particular additive abelian groups, φ is in particular a homomorphism of abelian groups and the fibers of φ are the additive cosets $r + I$ of the kernel I (more precisely,

if r is any element of R mapping to $a \in S$, $\varphi(r) = a$, then the fiber of φ over a is the coset $r + I$ of the kernel I). These fibers have the structure of a ring naturally isomorphic to the image of φ: if X is the fiber over $a \in S$ and Y is the fiber over $b \in S$, then $X + Y$ is the fiber over $a + b$ and XY is the fiber over ab. In terms of cosets of the kernel I this addition and multiplication is

$$(r + I) + (s + I) = (r + s) + I \tag{7.1}$$

$$(r + I) \times (s + I) = (rs) + I. \tag{7.2}$$

As in the case for groups, the verification that these operations define a ring structure on the collection of cosets of the kernel I ultimately rests on the corresponding ring properties of S. This ring of cosets is called the *quotient ring* of R by $I = \ker \varphi$ and is denoted R/I. Note that the additive structure of the ring R/I is just the additive quotient group of the additive abelian group R by the (necessarily normal) subgroup I. When I is the kernel of some homomorphism φ this additive abelian quotient group also has a multiplicative structure, defined by (7.2), which makes R/I into a ring.

As in the case for groups, we can also consider whether (7.1) and (7.2) can be used to define a ring structure on the collection of cosets of an *arbitrary* subgroup I of R. Note that since R is an abelian additive group, the subgroup I is necessarily normal so that the quotient R/I of cosets of I is automatically an additive abelian group. The question then is whether this quotient group also has a *multiplicative* structure induced from the multiplication in R, defined by (7.2). The answer is no in general (just as the answer is no in trying to form the quotient by an arbitrary subgroup of a group), which leads to the notion of an *ideal* in R (the analogue for rings of a normal subgroup of a group). We shall then see that the ideals of R are exactly the kernels of the ring homomorphisms of R (the analogue for rings of the characterization of normal subgroups as the kernels of group homomorphisms).

Let I be an arbitrary subgroup of the additive group R. We consider when the multiplication of cosets in (7.2) is well defined and makes the additive abelian group R/I into a ring. The statement that the multiplication in (7.2) is well defined is the statement that the multiplication is independent of the particular representatives r and s chosen, i.e. that we obtain the same coset on the right if instead we use the representatives $r + \alpha$ and $s + \beta$ for any $\alpha, \beta \in I$. In other words, we must have

$$(r + \alpha)(s + \beta) + I = rs + I \tag{$*$}$$

for all $r, s \in R$ and all $\alpha, \beta \in I$.

Letting $r = s = 0$, we see that I must be closed under multiplication, i.e. I must be a *subring* of R.

Next, by letting $s = 0$ and letting r be arbitrary, we see that we must have $r\beta \in I$ for every $r \in R$ and every $\beta \in I$, i.e. that I must be closed under multiplication on the left by elements from R. Letting $r = 0$ and letting s be arbitary, we see similarly that I must be closed under multiplication on the right by elements from R.

Conversely, if I is closed under multiplication on the left and on the right by elements from R then the relation $(*)$ is satisfied for all $\alpha, \beta \in I$. Hence this is a necessary and sufficient condition for the multiplication in (7.2) to be well defined.

Finally, if the multiplication of cosets defined by (7.2) is well defined, then this multiplication makes the additive quotient group R/I into a ring. Each ring axiom in the quotient follows directly from the corresponding axiom in R. For example, one of the distributive laws is verified as follows:

$$\begin{aligned}(r + I)[(s + I) + (t + I)] &= (r + I)[(s + t) + I] \\ &= r(s + t) + I = (rs + rt) + I \\ &= (rs + I) + (rt + I) \\ &= [(r + I)(s + I)] + [(r + I)(t + I)].\end{aligned}$$

This shows that the quotient R/I of the ring R by a subgroup I has a natural ring structure if and only if I is also closed under multiplication on the left and on the right by elements from R (so in particular must be a subring of R since it is closed under multiplication). As mentioned, such subrings I are called the *ideals* of R.

Definition. Let R be a ring, let I be a subset of R and let $r \in R$.

(1) $rI = \{ra \mid a \in I\}$ and $Ir = \{ar \mid a \in I\}$.

(2) A subset I of R is a *left ideal* of R if

- **(i)** I is a subring of R, and
- **(ii)** I is closed under left multiplication by elements from R, i.e. $rI \subseteq I$ for all $r \in R$.

Similarly I is a *right ideal* if (i) holds and in place of (ii) one has

- **(ii)′** I is closed under right multiplication by elements from R, i.e. $Ir \subseteq I$ for all $r \in R$.

(3) A subset I that is both a left ideal and a right ideal is called an *ideal* (or, for added emphasis, a *two-sided ideal*) of R.

For commutative rings the notions of left, right and two-sided ideal coincide. We emphasize that to prove a subset I of a ring R is an ideal it is necessary to prove that I is nonempty, closed under subtraction and closed under multiplication by all the elements of R (and not just by elements of I). If R has a 1 then $(-1)a = -a$ so in this case I is an ideal if it is nonempty, closed under addition and closed under multiplication by all the elements of R.

Note also that the last part of Proposition 5 proves that the kernel of any ring homomorphism is an ideal.

We summarize the preceding discussion in the following proposition.

Proposition 6. Let R be a ring and let I be an ideal of R. Then the (additive) quotient group R/I is a ring under the binary operations:

$$(r + I) + (s + I) = (r + s) + I \qquad \text{and} \qquad (r + I) \times (s + I) = (rs) + I$$

for all $r, s \in R$. Conversely, if I is any subgroup such that the above operations are well defined, then I is an ideal of R.

Definition. When I is an ideal of R the ring R/I with the operations in the previous proposition is called the *quotient ring* of R by I.

Theorem 7

(1) (The First Isomorphism Theorem) If $\varphi : R \to S$ is a homomorphism of rings, then the kernel of φ is an ideal of R, the image of φ is a subring of S and $R/\ker\varphi$ is isomorphic as a ring to $\varphi(R)$.

(2) If I is any ideal of R, then the map

$$R \to R/I \qquad \text{defined by} \qquad r \mapsto r + I$$

is a surjective ring homomorphism with kernel I (this homomorphism is called the *natural projection* of R onto R/I). Thus every ideal is the kernel of a ring homomorphism and vice versa.

Proof: This is just a matter of collecting previous calculations. If I is the kernel of φ, then the cosets (under addition) of I are precisely the fibers of φ. In particular, the cosets $r + I$, $s + I$ and $rs + I$ are the fibers of φ over $\varphi(r), \varphi(s)$ and $\varphi(rs)$, respectively. Since φ is a ring homomorphism $\varphi(r)\varphi(s) = \varphi(rs)$, hence $(r+I)(s+I) = rs+I$. Multiplication of cosets is well defined and so I is an ideal and R/I is a ring. The correspondence $r + I \mapsto \varphi(r)$ is a bijection between the rings R/I and $\varphi(R)$ which respects addition and multiplication, hence is a ring isomorphism.

If I is any ideal, then R/I is a ring (in particular is an abelian group) and the map $\pi : r \mapsto r + I$ is a group homomorphism with kernel I. It remains to check that π is a ring homomorphism. This is immediate from the definition of multiplication in R/I:

$$\pi : rs \mapsto rs + I = (r + I)(s + I) = \pi(r)\pi(s).$$

As with groups we shall often use the bar notation for reduction mod I: $\overline{r} = r+I$. With this notation the addition and multiplication in the quotient ring R/I become simply $\overline{r} + \overline{s} = \overline{r + s}$ and $\overline{r}\,\overline{s} = \overline{rs}$.

Examples

Let R be a ring.

(1) The subrings R and $\{0\}$ are ideals. An ideal I is *proper* if $I \neq R$. The ideal $\{0\}$ is called the *trivial ideal* and is denoted by 0.

(2) It is immediate that $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$ for any $n \in \mathbb{Z}$ and these are the only ideals of $\mathbb{Z}$ since in particular these are the only subgroups of $\mathbb{Z}$. The associated quotient ring is $\mathbb{Z}/n\mathbb{Z}$ (which explains the choice of notation and which we have now proved is a ring), introduced in Chapter 0. For example, if $n = 15$ then the elements of $\mathbb{Z}/15\mathbb{Z}$ are the cosets $\overline{0}, \overline{1}, \ldots, \overline{13}, \overline{14}$. To add (or multiply) in the quotient, simply choose any representatives for the two cosets, add (multiply, respectively) these representatives in the integers $\mathbb{Z}$, and take the corresponding coset containing this sum (product, respectively). For example, $\overline{7} + \overline{11} = \overline{18}$ and $\overline{18} = \overline{3}$, so $\overline{7} + \overline{11} = \overline{3}$ in $\mathbb{Z}/15\mathbb{Z}$. Similarly, $\overline{7}\,\overline{11} = \overline{77} = \overline{2}$ in $\mathbb{Z}/15\mathbb{Z}$. We could also express this by writing $7 + 11 \equiv 3 \bmod 15$, $7(11) \equiv 2 \bmod 15$.

The natural projection $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is called *reduction mod n* and will be discussed further at the end of these examples.

(3) Let $R = \mathbb{Z}[x]$ be the ring of polynomials in x with integer coefficients. Let I be the collection of polynomials whose terms are of degree at least 2 (i.e. having no terms of degree 0 or degree 1) together with the zero polynomial. Then I is an ideal: the sum of two such poly-

nomials again has terms of degree at least 2 and the product of a polynomial whose terms are of degree at least 2 with *any* polynomial again only has terms of degree at least 2. Two polynomials $p(x), q(x)$ are in the same coset of I if and only if they differ by a polynomial whose terms are of degree at least 2, i.e. if and only if $p(x)$ and $q(x)$ have the same constant and first degree terms. For example, the polynomials $3 + 5x + x^3 + x^5$ and $3 + 5x - x^4$ are in the same coset of I. It follows easily that a complete set of representatives for the quotient R/I is given by the polynomials $a + bx$ of degree at most 1.

Addition and multiplication in the quotient are again performed by representatives. For example,

$$(\overline{1+3x}) + (\overline{-4+5x}) = \overline{-3+8x}$$

and

$$(\overline{1+3x})(\overline{-4+5x}) = \overline{(-4-7x+15x^2)} = \overline{-4-7x}\,.$$

Note that in this quotient ring R/I we have $\overline{x}\,\overline{x} = \overline{x^2} = \overline{0}$, for example, so that R/I has zero divisors, even though $R = \mathbb{Z}[x]$ does not.

(4) Let A be a ring, let X be any nonempty set and let R be the ring of all functions from X to A. For each fixed $c \in X$ the map

$$E_c : R \to A \quad \text{defined by} \quad E_c(f) = f(c)$$

(called *evaluation at* c) is a ring homomorphism because the operations in R are pointwise addition and multiplication of functions. The kernel of E_c is $\{f \in R \mid f(c) = 0\}$ (the set of functions from X to A that vanish at c). Also, E_c is surjective: given any $a \in A$ the constant function $f(x) = a$ maps to a under evaluation at c. Thus $R/\ker E_c \cong A$.

Similarly, let X be the closed interval $[0,1]$ in $\mathbb{R}$ and let R be the ring of all continuous real valued functions on $[0,1]$. For each $c \in [0,1]$, evaluation at c is a surjective ring homomorphism (since R contains the constant functions) and so $R/\ker E_c \cong \mathbb{R}$. The kernel of E_c is the ideal of all continuous functions whose graph crosses the x-axis at c. More generally, the fiber of E_c above the real number y_0 is the set of all continuous functions that pass through the point (c, y_0).

(5) The map from the polynomial ring $R[x]$ to R defined by $p(x) \mapsto p(0)$ (evaluation at 0) is a ring homomorphism whose kernel is the set of all polynomials whose constant term is zero, i.e. $p(0) = 0$. We can compose this homomorphism with any homomorphism from R to another ring S to obtain a ring homomorphism from $R[x]$ to S. For example, let $R = \mathbb{Z}$ and consider the homomorphism $\mathbb{Z}[x] \to \mathbb{Z}/2\mathbb{Z}$ defined by $p(x) \mapsto p(0) \mapsto p(0) \bmod 2 \in \mathbb{Z}/2\mathbb{Z}$. The kernel of this composite map is $\{p(x) \in \mathbb{Z}[x] \mid p(0) \in 2\mathbb{Z}\}$, i.e. the set of all polynomials with integer coefficients whose constant term is even. The other fiber of this homomorphism is the coset of polynomials whose constant term is odd, as we determined earlier. Since the homomorphism is clearly surjective, the quotient ring is $\mathbb{Z}/2\mathbb{Z}$.

(6) Fix some $n \in \mathbb{Z}$ with $n \geq 2$ and consider the noncommutative ring $M_n(R)$. If J is any ideal of R then $M_n(J)$, the $n \times n$ matrices whose entries come from J, is a two-sided ideal of $M_n(R)$. This ideal is the kernel of the surjective homomorphism $M_n(R) \to M_n(R/J)$ which reduces each entry of a matrix mod J, i.e. which maps each entry a_{ij} to $\overline{a_{ij}}$ (here bar denotes passage to R/J). For instance, when $n = 3$ and $R = \mathbb{Z}$, the 3×3 matrices whose entries are all even is the two-sided ideal $M_3(2\mathbb{Z})$ of $M_3(\mathbb{Z})$ and the quotient $M_3(\mathbb{Z})/M_3(2\mathbb{Z})$ is isomorphic to $M_3(\mathbb{Z}/2\mathbb{Z})$.

If the ring R has an identity then the exercises below show that every two-sided ideal of $M_n(R)$ is of the form $M_n(J)$ for some two-sided ideal J of R.

(7) Let R be a commutative ring with 1 and let $G = \{g_1, \ldots, g_n\}$ be a finite group. The map from the group ring RG to R defined by $\sum_{i=1}^n a_i g_i \mapsto \sum_{i=1}^n a_i$ is easily seen to be a homomorphism, called the *augmentation map*. The kernel of the augmentation map, the *augmentation ideal*, is the set of elements of RG whose coefficients sum to 0. For example, $g_i - g_j$ is an element of the augmentation ideal for all i, j. Since the augmentation map is surjective, the quotient ring is isomorphic to R.

Another ideal in RG is $\{\sum_{i=1}^n ag_i \mid a \in R\}$, i.e. the formal sums whose coefficients are all equal (equivalently, all R-multiples of the element $g_1 + \cdots + g_n$).

(8) Let R be a commutative ring with identity $1 \neq 0$ and let $n \in \mathbb{Z}$ with $n \geq 2$. We exhibit some one-sided ideals in the ring $M_n(R)$. For each $j \in \{1, 2, \ldots, n\}$ let L_j be the set of all $n \times n$ matrices in $M_n(R)$ with arbitrary entries in the jth column and zeros in all other columns. It is clear that L_j is closed under subtraction. It follows directly from the definition of matrix multiplication that for any matrix $T \in M_n(R)$ and any $A \in L_j$ the product TA has zero entries in the ith column for all $i \neq j$. This shows L_j is a *left ideal* of $M_n(R)$. Moreover, L_j is *not* a *right* ideal (hence is not a two-sided ideal). To see this, let E_{pq} be the matrix with 1 in the pth row and q^{th} column and zeros elsewhere ($p, q \in \{1, \ldots, n\}$). Then $E_{1j} \in L_j$ but $E_{1j}E_{ji} = E_{1i} \notin L_j$ if $i \neq j$, so L_j is not closed under right multiplication by arbitrary ring elements. An analogous argument shows that if R_j is the set of all $n \times n$ matrices in $M_n(R)$ with arbitrary entries in the jth row and zeros in all other rows, then R_j is a *right* ideal which is not a *left* ideal. These one-sided ideals will play an important role in Part V.

Example: (The Reduction Homomorphism)

The canonical projection map from $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ obtained by factoring out by the ideal $n\mathbb{Z}$ of $\mathbb{Z}$ is usually referred to as "reducing modulo n". The fact that this is a *ring homomorphism* has important consequences for elementary number theory. For example, suppose we are trying to solve the equation

$$x^2 + y^2 = 3z^2$$

in *integers* x, y and z (such problems are frequently referred to as *Diophantine equations* after Diophantus, who was one of the first to systematically examine the existence of *integer* solutions of equations). Suppose such integers exist. Observe first that we may assume x, y and z have no factors in common, since otherwise we could divide through this equation by the square of this common factor and obtain another set of integer solutions smaller than the initial ones. This equation simply states a relation between these elements in the *ring* $\mathbb{Z}$. As such, the same relation must also hold in any *quotient* ring as well. In particular, this relation must hold in $\mathbb{Z}/n\mathbb{Z}$ for any integer n. The choice $n = 4$ is particularly efficacious, for the following reason: the squares mod 4 are just $0^2, 1^2, 2^2, 3^2$, i.e. $0, 1 \pmod 4$. Reading the above equation mod 4 (that is, considering this equation in the quotient ring $\mathbb{Z}/4\mathbb{Z}$), we must have

$$\left\{\begin{matrix} 0 \\ 1 \end{matrix}\right\} + \left\{\begin{matrix} 0 \\ 1 \end{matrix}\right\} \equiv 3\left\{\begin{matrix} 0 \\ 1 \end{matrix}\right\} \equiv \left\{\begin{matrix} 0 \\ 3 \end{matrix}\right\} \pmod 4$$

where the $\left\{\begin{matrix} 0 \\ 1 \end{matrix}\right\}$, for example, indicates that either a 0 or a 1 may be taken. Checking the few possibilities shows that we must take the 0 each time. This means that each of

x, y and z must be even integers (squares of the odd integers gave us 1 mod 4). But this contradicts the assumption of no common factors for these integers, and shows that this equation has *no solutions in nonzero integers*.

Note that even had solutions existed, this technique gives information about the possible residues of the solutions mod n (since we could just as well have examined the possibilities mod n as mod 4) and note that for each choice of n we have only a *finite* problem to solve because there are only finitely many residue classes mod n. Together with the Chinese Remainder Theorem (described in Section 6), we can then determine the possible solutions modulo very large integers, which greatly assists in finding them numerically (when they exist). We also observe that this technique has a number of limitations — for example, there are equations which have solutions modulo every integer, but which do not have integer solutions. An easy example (but extremely hard to verify that it does indeed have this property) is the equation

$$3x^3 + 4y^3 + 5z^3 = 0.$$

As a final example of this technique, we mention that the map from the ring $\mathbb{Z}[x]$ of polynomials with integer coefficients to the ring $\mathbb{Z}/p\mathbb{Z}[x]$ of polynomials with coefficients in $\mathbb{Z}/p\mathbb{Z}$ for a prime p given by *reducing the coefficients modulo p* is a ring homomorphism. This example of reduction will be used in Chapter 9 in trying to determine whether polynomials can be factored.

We close this section with the remaining Isomorphism Theorems for rings. Each of these may be proved ~~as follows: first~~ use the corresponding theorem from group theory to obtain an isomorphism of *additive groups* (or correspondence of groups, in the case of the Fourth Isomorphism Theorem) and then check that this group isomorphism (or correspondence, respectively) is a multiplicative map, and so defines a *ring* isomorphism. In each case the verification is immediate from the definition of multiplication in quotient rings. For example, the map that gives the isomorphism in (2) below is defined by $\varphi : r + I \mapsto r + J$. This map is multiplicative since $(r_1 + I)(r_2 + I) = r_1r_2 + I$ by the definition of the multiplication in the quotient ring R/I, and $r_1r_2 + I \mapsto r_1r_2 + J = (r_1 + J)(r_2 + J)$ by the definition of the multiplication in the quotient ring R/J, i.e. $\varphi(r_1r_2) = \varphi(r_1)\varphi(r_2)$. The proofs for the other parts of the theorem are similar.

Theorem 8. Let R be a ring.

(1) (The Second Isomorphism Theorem for Rings) Let A be a subring and let B be an ideal of R. Then $A + B = \{a + b \mid a \in A,\ b \in B\}$ is a subring of R, $A \cap B$ is an ideal of A and $(A + B)/B \cong A/(A \cap B)$.

(2) (The Third Isomorphism Theorem for Rings) Let I and J be ideals of R with $I \subseteq J$. Then J/I is an ideal of R/I and $(R/I)/(J/I) \cong R/J$.

(3) (The Fourth or Lattice Isomorphism Theorem for Rings) Let I be an ideal of R. The correspondence $A \leftrightarrow A/I$ is an inclusion preserving bijection between the set of subrings A of R that contain I and the set of subrings of R/I. Furthermore, A (a subring containing I) is an ideal of R if and only if A/I is an ideal of R/I.

Example

Let $R = \mathbb{Z}$ and let I be the ideal $12\mathbb{Z}$. The quotient ring $\overline{R} = R/I = \mathbb{Z}/12\mathbb{Z}$ has ideals $\overline{R}$, $2\mathbb{Z}/12\mathbb{Z}$, $3\mathbb{Z}/12\mathbb{Z}$, $4\mathbb{Z}/12\mathbb{Z}$, $6\mathbb{Z}/12\mathbb{Z}$, and $\overline{0} = 12\mathbb{Z}/12\mathbb{Z}$ corresponding to the ideals $R = \mathbb{Z}, 2\mathbb{Z}, 3\mathbb{Z}, 4\mathbb{Z}, 6\mathbb{Z}$ and $12\mathbb{Z} = I$ of R containing I, respectively.

EXERCISES

Let R be a ring with identity $1 \neq 0$.

1. Prove that the rings $2\mathbb{Z}$ and $3\mathbb{Z}$ are not isomorphic.

2. Prove that the rings $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ are not isomorphic.

3. Find all homomorphic images of $\mathbb{Z}$.

4. Find all ring homomorphisms from $\mathbb{Z}$ to $\mathbb{Z}/30\mathbb{Z}$. In each case describe the kernel and the image.

5. Describe all ring homomorphisms from the ring $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z}$. In each case describe the kernel and the image.

6. Decide which of the following are ring homomorphisms from $M_2(\mathbb{Z})$ to $\mathbb{Z}$:

(a) $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto a$ (projection onto the 1,1 entry)

(b) $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto a + d$ (the *trace* of the matrix)

(c) $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto ad - bc$ (the *determinant* of the matrix).

7. Let $R = \{\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mid a, b, d \in \mathbb{Z}\}$ be the subring of $M_2(\mathbb{Z})$ of upper triangular matrices. Prove that the map

$$\varphi : R \to \mathbb{Z} \times \mathbb{Z} \quad \text{defined by} \quad \varphi : \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mapsto (a, d)$$

is a surjective homomorphism and describe its kernel.

8. Decide which of the following are ideals of the ring $\mathbb{Z} \times \mathbb{Z}$:

(a) $\{(a, a) \mid a \in \mathbb{Z}\}$
(b) $\{(2a, 2b) \mid a, b \in \mathbb{Z}\}$
(c) $\{(2a, 0) \mid a \in \mathbb{Z}\}$
(d) $\{(a, -a) \mid a \in \mathbb{Z}\}$.

9. Decide which of the sets in Exercise 6 of Section 1 are ideals of the ring of all functions from [0,1] to $\mathbb{R}$.

10. Decide which of the following are ideals of the ring $\mathbb{Z}[x]$:

(a) the set of all polynomials whose constant term is a multiple of 3
(b) the set of all polynomials whose coefficient of x^2 is a multiple of 3
(c) the set of all polynomials whose constant term, coefficient of x and coefficient of x^2 are zero
(d) $\mathbb{Z}[x^2]$ (i.e. the polynomials in which only even powers of x appear)
(e) the set of polynomials whose coefficients sum to zero
(f) the set of polynomials $p(x)$ such that $p'(0) = 0$, where $p'(x)$ is the usual first derivative of $p(x)$ with respect to x.

11. Let R be the ring of all continuous real valued functions on the closed interval $[0,1]$. Prove that the map $\varphi : R \to \mathbb{R}$ defined by $\varphi(f) = \int_0^1 f(t)dt$ is a homomorphism of additive groups but not a ring homomorphism.

12. Let D be an integer that is not a perfect square in $\mathbb{Z}$ and let $S = \{\begin{pmatrix} a & b \\ Db & a \end{pmatrix} \mid a, b \in \mathbb{Z}\}$.
    (a) Prove that S is a subring of $M_2(\mathbb{Z})$.
    (b) Prove that the map $\varphi : \mathbb{Z}[\sqrt{D}] \to S$ defined by $\varphi(a + b\sqrt{D}) = \begin{pmatrix} a & b \\ Db & a \end{pmatrix}$ is a ring isomorphism (the ring $\mathbb{Z}[\sqrt{D}]$ is defined in Exercise 23, Section 1).

13. Prove that the ring $M_2(\mathbb{R})$ contains a subring that is isomorphic to $\mathbb{C}$.

14. Prove that the ring $M_4(\mathbb{R})$ contains a subring that is isomorphic to the real Hamilton quaternions, $\mathbb{H}$.

15. Let X be a nonempty set and let $\mathcal{P}(X)$ be the Boolean ring of all subsets of X defined in Exercise 21 of Section 1. Let R be the ring of all functions from X into $\mathbb{Z}/2\mathbb{Z}$. For each $A \in \mathcal{P}(X)$ define the function

$$\chi_A : X \to \mathbb{Z}/2\mathbb{Z} \qquad \text{by} \qquad \chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

(χ_A is called the *characteristic function of* A with values in $\mathbb{Z}/2\mathbb{Z}$). Prove that the map $\mathcal{P}(X) \to R$ defined by $A \mapsto \chi_A$ is a ring isomorphism.

16. Let $\varphi : R \to S$ be a surjective homomorphism of rings. Prove that the image of the center of R is contained in the center of S (cf. Exercise 7 of Section 1).

17. Let R and S be nonzero rings with identity and denote their respective identities by 1_R and 1_S. Let $\varphi : R \to S$ be a homomorphism of rings.
    (a) Prove that if $\varphi(1_R) \neq 1_S$ then $\varphi(1_R)$ is a zero divisor in S. Deduce that if S is an integral domain then every ring homomorphism from R to S sends the identity of R to the identity of S.
    (b) Prove that if $\varphi(1_R) = 1_S$ then $\varphi(u)$ is a unit in S for each unit u of R and $\varphi(u^{-1}) = \varphi(u)^{-1}$.

18. (a) If I and J are ideals of R prove that their intersection $I \cap J$ is also an ideal of R.
    (b) Prove that the intersection of an arbitrary nonempty collection of ideals is again an ideal (do not assume the collection is countable).

19. Prove that if $I_1 \subseteq I_2 \subseteq \cdots$ are ideals of R then $\cup_{n=1}^{\infty} I_n$ is an ideal of R.

20. Let I be an ideal of R and let S be a subring of R. Prove that $I \cap S$ is an ideal of S. Show by example that not every ideal of a subring S of a ring R need be of the form $I \cap S$ for some ideal I of R.

21. Prove that every (two-sided) ideal of $M_n(R)$ is equal to $M_n(J)$ for some (two-sided) ideal J of R. [Use Exercise 6(c) of Section 2 to show first that the set of entries of matrices in an ideal of $M_n(R)$ form an ideal in R.]

22. Let a be an element of the ring R.
    (a) Prove that $\{x \in R \mid ax = 0\}$ is a right ideal and $\{y \in R \mid ya = 0\}$ is a left ideal (called respectively the right and left *annihilators* of a in R).
    (b) Prove that if L is a left ideal of R then $\{x \in R \mid xa = 0 \text{ for all } a \in L\}$ is a two-sided ideal (called the left *annihilator* of L in R).

23. Let S be a subring of R and let I be an ideal of R. Prove that if $S \cap I = 0$ then $\overline{S} \cong S$, where the bar denotes passage to R/I.

24. Let $\varphi : R \to S$ be a ring homomorphism.

(a) Prove that if J is an ideal of S then $\varphi^{-1}(J)$ is an ideal of R. Apply this to the special case when R is a subring of S and φ is the inclusion homomorphism to deduce that if J is an ideal of S then $J \cap R$ is an ideal of R.

(b) Prove that if φ is surjective and I is an ideal of R then $\varphi(I)$ is an ideal of S. Give an example where this fails if φ is not surjective.

25. Assume R is a commutative ring with 1. Prove that the Binomial Theorem

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

holds in R, where the binomial coefficient $\binom{n}{k}$ is interpreted in R as $1 + 1 + \cdots + 1$ ($\binom{n}{k}$ times).

26. The *characteristic* of a ring R is the smallest positive integer n such that $1 + 1 + \cdots + 1 = 0$ (n times) in R; if no such integer exists the characteristic of R is said to be 0. For example, $\mathbb{Z}/n\mathbb{Z}$ is a ring of characteristic n for each positive integer n and $\mathbb{Z}$ is a ring of characteristic 0.

(a) Prove that the map $\mathbb{Z} \to R$ defined by

$$k \mapsto \begin{cases} 1 + 1 + \cdots + 1 \ (k \text{ times}) & \text{if } k > 0 \\ 0 & \text{if } k = 0 \\ -1 - 1 - \cdots - 1 \ (-k \text{ times}) & \text{if } k < 0 \end{cases}$$

is a ring homomorphism whose kernel is $n\mathbb{Z}$, where n is the characteristic of R (this explains the use of the terminology "characteristic 0" instead of the archaic phrase "characteristic ∞" for rings in which no sum of 1's is zero).

(b) Determine the characteristics of the rings: $\mathbb{Q}$, $\mathbb{Z}[x]$, $\mathbb{Z}/n\mathbb{Z}[x]$.

(c) Prove that if p is a prime and if R is a commutative ring of characteristic p, then $(a+b)^p = a^p + b^p$ for all $a, b \in R$.

27. Prove that a nonzero Boolean ring has characteristic 2 (cf. Exercise 15, Section 1).

28. Prove that an integral domain has characteristic p, where p is either a prime or 0 (cf. Exercise 26).

29. Let R be a commutative ring. Recall (cf. Exercise 13, Section 1) that an element $x \in R$ is nilpotent if $x^n = 0$ for some $n \in \mathbb{Z}^+$. Prove that the set of nilpotent elements form an ideal — called the *nilradical* of R and denoted by $\eta(R)$. [Use the Binomial Theorem to show $\eta(R)$ is closed under addition.]

30. Prove that if R is a commutative ring and $\eta(R)$ is its nilradical (cf. the preceding problem) then zero is the only nilpotent element of $R/\eta(R)$ i.e. prove that $\eta(R/\eta(R)) = 0$.

31. Prove that the elements $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ are nilpotent elements of $M_2(\mathbb{Z})$ whose sum is not nilpotent (note that these two matrices do not commute). Deduce that the set of nilpotent elements in the noncommutative ring $M_2(\mathbb{Z})$ is not an ideal.

32. Let $\varphi : R \to S$ be a homomorphism of rings. Prove that if x is a nilpotent element of R then $\varphi(x)$ is nilpotent in S.

33. Assume R is commutative. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be an element of the polynomial ring $R[x]$.

(a) Prove that $p(x)$ is a unit in $R[x]$ if and only if a_0 is a unit and $a_1, a_2, \ldots, a_n$ are nilpotent in R. [See Exercise 14 of Section 1.]

(b) Prove that $p(x)$ is nilpotent in $R[x]$ if and only if $a_0, a_1, \ldots, a_n$ are nilpotent elements of R.

34. Let I and J be ideals of R.

(a) Define $I + J = \{a + b \mid a \in I,\ b \in J\}$. Prove that $I + J$ is an ideal and that it is the smallest ideal of R containing both I and J.

(b) Define IJ to be the set of all finite sums of elements of the form ab with $a \in I$ and $b \in J$. Prove that IJ is an ideal contained in $I \cap J$.

(c) Give an example where $IJ \neq I \cap J$.

(d) Prove that if R is commutative and if $I + J = R$ then $IJ = I \cap J$.

35. Let I, J, K be ideals of R.

(a) Prove that $I(J + K) = IJ + IK$ and $(I + J)K = IK + JK$.

(b) Prove that if $J \subseteq I$ then $I \cap (J + K) = J + (I \cap K)$.

36. If I is an ideal of any ring R and n is a positive integer let I^n be the ideal consisting of all finite sums of elements of the form $a_1 a_2 \cdots a_n$ with $a_i \in I$ for all i. Equivalently, I^n may be defined inductively as II^{n-1}, where the product of two ideals is defined in Exercise 34. Note that I^n is not just sums of nth powers of elements of I. Show that if I is the ideal of all polynomials in $\mathbb{Z}[x]$ with zero constant term then $I^n = \{a_n x^n + a_{n+1} x^{n+1} + \cdots + a_{n+m} x^{n+m} \mid a_i \in \mathbb{Z},\ m \geq 0\}$.

37. An ideal N is called *nilpotent* if N^n is the zero ideal for some $n \in \mathbb{Z}^+$ (cf. the preceding exercise). Prove that the ideal $p\mathbb{Z}/p^m\mathbb{Z}$ is a nilpotent ideal in the ring $\mathbb{Z}/p^m\mathbb{Z}$.

7.4 PROPERTIES OF IDEALS

Throughout this section R is a ring with identity $1 \neq 0$.

Definition. Let A be any subset of the ring R.

(1) Let (A) denote the smallest ideal of R containing A, called *the ideal generated by* A.

(2) Let RA denote the set of all finite sums of elements of the form ra with $r \in R$ and $a \in A$ i.e. $RA = \{r_1 a_1 + r_2 a_2 + \cdots + r_n a_n \mid r_i \in R,\ a_i \in A,\ n \in \mathbb{Z}^+\}$ (where the convention is $RA = 0$ if $A = \emptyset$).
Similarly, $AR = \{a_1 r_1 + a_2 r_2 + \cdots + a_n r_n \mid r_i \in R,\ a_i \in A,\ n \in \mathbb{Z}^+\}$ and $RAR = \{r_1 a_1 r_1' + r_2 a_2 r_2' + \cdots + r_n a_n r_n' \mid r_i, r_i' \in R,\ a_i \in A,\ n \in \mathbb{Z}^+\}$.

(3) An ideal generated by a single element is called a *principal ideal*.

(4) An ideal generated by a finite set is called a *finitely generated ideal*.

When $A = \{a\}$ or $\{a_1, a_2, \ldots\}$, etc., we shall drop the set brackets and simply write (a), $(a_1, a_2, \ldots)$ for (A), respectively.

The notion of ideals generated by subsets of a ring is analogous to that of subgroups generated by subsets of a group (Section 2.4). Since the intersection of any nonempty collection of ideals of R is also an ideal (cf. Exercise 18, Section 3) and A is always contained in at least one ideal (namely R), we have

$$(A) = \bigcap_{\substack{I \text{ an ideal} \\ A \subseteq I}} I,$$

i.e. (A) is the intersection of all ideals of R that contain the set A.

The *left ideal generated by* A is the intersection of all left ideals of R that contain A. This left ideal is obtained from A by closing A under all the operations that define a left ideal. It is immediate from the definition that RA is closed under addition and under left multiplication by any ring element. Since R has an identity, RA contains A. Thus RA is a left ideal of R which contains A. Conversely, any left ideal which contains A must contain all finite sums of elements of the form ra, $r \in R$ and $a \in A$ and so must contain RA. Thus *RA is precisely the left ideal generated by A*. Similarly, *AR is the right ideal generated by A* and *RAR is the (two-sided) ideal generated by A*. In particular,

$$if\ R\ is\ commutative\ then\ RA = AR = RAR = (A).$$

When R is a commutative ring and $a \in R$, the principal ideal (a) generated by a is just the set of all R-multiples of a. If R is not commutative, however, the set $\{ras \mid r, s \in R\}$ is not necessarily the two-sided ideal generated by a since it need not be closed under addition (in this case the ideal generated by a is the ideal RaR, which consists of all *finite sums* of elements of the form ras, $r, s \in R$).

The formation of principal ideals in a commutative ring is a particularly simple way of creating ideals, similar to generating cyclic subgroups of a group. Notice that the element $b \in R$ belongs to the ideal (a) if and only if $b = ra$ for some $r \in R$, i.e. if and only if *b is a multiple of a* or, put another way, *a divides b in R*. Also, $b \in (a)$ if and only if $(b) \subseteq (a)$. Thus containment relations between ideals, in particular between principal ideals, is seen to capture some of the arithmetic of general commutative rings. Commutative rings in which all ideals are principal are among the easiest to study and these will play an important role in Chapters 8 and 9.

Examples

(1) The trivial ideal 0 and the ideal R are both principal: $0 = (0)$ and $R = (1)$.

(2) In $\mathbb{Z}$ we have $n\mathbb{Z} = \mathbb{Z}n = (n) = (-n)$ for all integers n. Thus our notation for aR is consistent with the definition of $n\mathbb{Z}$ we have been using. As noted in the preceding section, these are all the ideals of $\mathbb{Z}$ so *every* ideal of $\mathbb{Z}$ is principal. For positive integers n and m, $n\mathbb{Z} \subseteq m\mathbb{Z}$ if and only if m divides n in $\mathbb{Z}$, so the lattice of ideals containing $n\mathbb{Z}$ is the same as the lattice of divisors of n. Furthermore, the ideal generated by two nonzero integers n and m is the principal ideal generated by their greatest common divisor, d: $(n, m) = (d)$. The notation for (n, m) as the greatest common divisor of n and m is thus consistent with the same notation for the ideal generated by n and m (although a principal generator for the ideal generated by n and m is only determined up to a $\pm$ sign — we could make it unique by choosing a nonnegative generator). In particular, n and m are relatively prime if and only if $(n, m) = (1)$.

(3) We show that the ideal $(2, x)$ generated by 2 and x in $\mathbb{Z}[x]$ is *not* a principal ideal. Observe that $(2, x) = \{2p(x) + xq(x) \mid p(x), q(x) \in \mathbb{Z}[x]\}$ and so this ideal consists precisely of the polynomials with integer coefficients whose constant term is even (as discussed in Example 5 in the preceding section) — in particular, this is a proper ideal. Assume by way of contradiction that $(2, x) = (a(x))$ for some $a(x) \in \mathbb{Z}[x]$. Since $2 \in (a(x))$ there must be some $p(x)$ such that $2 = p(x)a(x)$. The degree of $p(x)a(x)$ equals degree $p(x)$ + degree $a(x)$, hence both $p(x)$ and $a(x)$ must be constant polynomials, i e. integers. Since 2 is a prime number, $a(x),\ p(x) \in \{\pm 1,\ \pm 2\}$. If $a(x)$ were ± 1 then every polynomial would be a multiple of $a(x)$, contrary to $(a(x))$ being

a proper ideal. The only possibility is $a(x) = \pm 2$. But now $x \in (a(x)) = (2) = (-2)$ and so $x = 2q(x)$ for some polynomial $q(x)$ with integer coefficients, clearly impossible. This contradiction proves that $(2, x)$ is not principal.

Note that the symbol (A) is ambiguous if the ring is not specified: the ideal generated by 2 and x in $\mathbb{Q}[x]$ is the entire ring (1) since it contains the element $\frac{1}{2}2 = 1$.

We shall see in Chapter 9 that for any *field* F, all ideals of $F[x]$ *are* principal.

(4) If R is the ring of all functions from the closed interval $[0,1]$ into $\mathbb{R}$ let M be the ideal $\{f \mid f(\frac{1}{2}) = 0\}$ (the kernel of evaluation at $\frac{1}{2}$). Let $g(x)$ be the function which is zero at $x = \frac{1}{2}$ and 1 at all other points. Then $f = fg$ for all $f \in M$ so M is a principal ideal with generator g. In fact, *any* function which is zero at $\frac{1}{2}$ and nonzero at all other points is another generator for the same ideal M.

On the other hand, if R is the ring of all *continuous* functions from $[0,1]$ to $\mathbb{R}$ then $\{f \mid f(\frac{1}{2}) = 0\}$ is *not* principal nor is it even finitely generated (cf. the exercises).

(5) If G is a finite group and R is a commutative ring with 1 then the augmentation ideal is generated by the set $\{g - 1 \mid g \in G\}$, although this need not be a minimal set of generators. For example, if G is a cyclic group with generator σ, then the augmentation ideal is a principal ideal with generator $\sigma - 1$.

Proposition 9. Let I be an ideal of R.

(1) $I = R$ if and only if I contains a unit.

(2) Assume R is commutative. Then R is a field if and only if its only ideals are 0 and R.

Proof: (1) If $I = R$ then I contains the unit 1. Conversely, if u is a unit in I with inverse v, then for any $r \in R$

$$r = r \cdot 1 = r(vu) = (rv)u \in I$$

hence $R = I$.

(2) The ring R is a field if and only if every nonzero element is a unit. If R is a field every nonzero ideal contains a unit, so by the first part R is the only nonzero ideal. Conversely, if 0 and R are the only ideals of R let u be any nonzero element of R. By hypothesis $(u) = R$ and so $1 \in (u)$. Thus there is some $v \in R$ such that $1 = vu$, i.e. u is a unit. Every nonzero element of R is therefore a unit and so R is a field.

Corollary 10. If R is a field then any nonzero ring homomorphism from R into another ring is an injection.

Proof: The kernel of a ring homomorphism is an ideal. The kernel of a nonzero homomorphism is a proper ideal hence is 0 by the proposition.

These results show that the ideal structure of fields is trivial. Our approach to studying an algebraic structure through its homomorphisms will still play a fundamental role in field theory (Part IV) when we study injective homomorphisms (embeddings) of one field into another and automorphisms of fields (isomorphisms of a field to itself).

If D is a ring with identity $1 \neq 0$ in which the only left ideals and the only right ideals are 0 and D, then D is a division ring. Conversely, the only (left, right or two-

sided) ideals in a division ring D are 0 and D, which gives an analogue of Proposition 9(2) if R is not commutative (see the exercises). However, if F is a field, then for any $n \geq 2$ the only two-sided ideals in the matrix ring $M_n(F)$ are 0 and $M_n(F)$, even though this is not a division ring (it does have proper, nontrivial, left and right ideals: cf. Section 3), which shows that Proposition 9(2) does not hold for noncommutative rings. Rings whose only two-sided ideals are 0 and the whole ring (which are called *simple* rings) will be studied in Chapter 15.

One important class of ideals are those which are not contained in any other proper ideal:

Definition. An ideal M in an arbitrary ring S is called a *maximal ideal* if $M \neq S$ and the only ideals containing M are M and S.

A general ring need not have maximal ideals. For example, take any abelian group which has no maximal subgroups (for example, $\mathbb{Q}$ – cf. Exercise 16, Section 6.1) and make it into a trivial ring by defining $ab = 0$ for all a, b. In such a ring the ideals are simply the subgroups and so there are no maximal ideals. The zero ring has no maximal ideals, hence any result involving maximal ideals forces a ring to be nonzero. The next proposition shows that rings with an identity $1 \neq 0$ always possess maximal ideals. Like many such general existence theorems (e.g. the result that a finitely generated group has maximal subgroups or that every vector space has a basis) the proof relies on Zorn's Lemma (see the Appendix). In many specific rings, however, the presence of maximal ideals is often obvious, independent of Zorn's Lemma.

Proposition 11. In a ring with identity every proper ideal is contained in a maximal ideal.

Proof: Let R be a ring with identity and let I be a proper ideal (so R cannot be the zero ring, i.e. $1 \neq 0$). Let $\mathcal{S}$ be the set of all proper ideals of R which contain I. Then $\mathcal{S}$ is nonempty ($I \in \mathcal{S}$) and is partially ordered by inclusion. If $\mathcal{C}$ is a chain in $\mathcal{S}$, define J to be the union of all ideals in $\mathcal{C}$:

$$J = \bigcup_{A \in \mathcal{C}} A.$$

We first show that J is an ideal. Certainly J is nonempty because $\mathcal{C}$ is nonempty – specifically, $0 \in J$ since 0 is in every ideal A. If $a, b \in J$, then there are ideals $A, B \in \mathcal{C}$ such that $a \in A$ and $b \in B$. By definition of a chain either $A \subseteq B$ or $B \subseteq A$. In either case $a - b \in J$, so J is closed under subtraction. Since each $A \in \mathcal{C}$ is closed under left and right multiplication by elements of R, so is J. This proves J is an ideal.

If J is not a proper ideal then $1 \in J$. In this case, by definition of J we must have $1 \in A$ for some $A \in \mathcal{C}$. This is a contradiction because each A is a proper ideal ($A \in \mathcal{C} \subseteq \mathcal{S}$). This proves that each chain has an upper bound in $\mathcal{S}$. By Zorn's Lemma $\mathcal{S}$ has a maximal element which is therefore a maximal (proper) ideal containing I.

For commutative rings the next result characterizes maximal ideals by the structure of their quotient rings.

Proposition 12. Assume R is commutative. The ideal M is a maximal ideal if and only if the quotient ring R/M is a field.

Proof: This follows from the Lattice Isomorphism Theorem together with Proposition 9(2). The ideal M is maximal if and only if there are no ideals I with $M \subset I \subset R$. By the Lattice Isomorphism Theorem the ideals of R containing M correspond bijectively with the ideals of R/M, so M is maximal if and only if the only ideals of R/M are 0 and R/M. By Proposition 9(2) we see that M is maximal if and only if R/M is a field.

The proposition above indicates how to *construct* some fields: take the quotient of any commutative ring R with identity by a maximal ideal in R. We shall use this in Part IV to construct all finite fields by taking quotients of the ring $\mathbb{Z}[x]$ by maximal ideals.

Examples

(1) Let n be a nonnegative integer. The ideal $n\mathbb{Z}$ of $\mathbb{Z}$ is a maximal ideal if and only if $\mathbb{Z}/n\mathbb{Z}$ is a field. We saw in Section 3 that this is the case if and only if n is a prime number. This also follows directly from the containment of ideals of $\mathbb{Z}$ described in Example 2 above.

(2) The ideal $(2, x)$ is a maximal ideal in $\mathbb{Z}[x]$ because its quotient ring is the field $\mathbb{Z}/2\mathbb{Z}$ — cf. Example 4 above and Example 5 at the end of Section 3.

(3) The ideal (x) in $\mathbb{Z}[x]$ is not a maximal ideal because $(x) \subset (2, x) \subset \mathbb{Z}[x]$. The quotient ring $\mathbb{Z}[x]/(x)$ is isomorphic to $\mathbb{Z}$ (the ideal (x) in $\mathbb{Z}[x]$ is the kernel of the surjective ring homomorphism from $\mathbb{Z}[x]$ to $\mathbb{Z}$ given by evaluation at 0). Since $\mathbb{Z}$ is not a field, we see again that (x) is not a maximal ideal in $\mathbb{Z}[x]$.

(4) Let R be the ring of all functions from [0,1] to $\mathbb{R}$ and for each $a \in [0, 1]$ let M_a be the kernel of evaluation at a. Since evaluation is a surjective homomorphism from R to $\mathbb{R}$, we see that $R/M_a \cong \mathbb{R}$ and hence M_a is a maximal ideal. Similarly, the kernel of evaluation at any fixed point is a maximal ideal in the ring of continuous real valued functions on $[0, 1]$.

(5) If F is a field and G is a finite group, then the augmentation ideal I is a maximal ideal of the group ring FG (cf. Example 7 at the end of the preceding section). The augmentation ideal is the kernel of the augmentation map which is a surjective homomorphism onto the field F (i.e. $FG/I \cong F$, a field). Note that Proposition 12 does not apply directly since FG need not be commutative, however, the implication in Proposition 12 that I is a maximal ideal if R/I is a field holds for arbitrary rings.

Definition. Assume R is commutative. An ideal P is called a *prime ideal* if $P \neq R$ and whenever the product ab of two elements $a, b \in R$ is an element of P, then at least one of a and b is an element of P.

The notion of a maximal ideal is fairly intuitive but the definition of a prime ideal may seem a little strange. It is, however, a natural generalization of the notion of a "prime" in the integers $\mathbb{Z}$. Let n be a nonnegative integer. According to the above definition the ideal $n\mathbb{Z}$ is a *prime* ideal provided $n \neq 1$ (to insure that the ideal is proper) and provided every time the product ab of two integers is an element of $n\mathbb{Z}$, at least one of a, b is an element of $n\mathbb{Z}$. Put another way, if $n \neq 0$, it must have the

property that whenever n divides ab, n must divide a or divide b. This is equivalent to the usual definition that n is a prime number. Thus *the prime ideals of $\mathbb{Z}$ are just the ideals $p\mathbb{Z}$ of $\mathbb{Z}$ generated by prime numbers p together with the ideal 0.*

For the integers $\mathbb{Z}$ there is no difference between the maximal ideals and the nonzero prime ideals. This is not true in general, but we shall see shortly that every maximal ideal is a prime ideal. First we translate the notion of prime ideals into properties of quotient rings as we did for maximal ideals in Proposition 12. Recall that an integral domain is a commutative ring with identity $1 \neq 0$ that has no zero divisors.

Proposition 13. Assume R is commutative. Then the ideal P is a prime ideal in R if and only if the quotient ring R/P is an integral domain.

Proof: This proof is simply a matter of translating the definition of a prime ideal into the language of quotients. The ideal P is prime if and only if $P \neq R$ and whenever $ab \in P$, then either $a \in P$ or $b \in P$. Use the bar notation for elements of R/P: $\bar{r} = r + P$. Note that $r \in P$ if and only if the element $\bar{r}$ is zero in the quotient ring R/P. Thus in the terminology of quotients P is a prime ideal if and only if $\overline{R} \neq \bar{0}$ and whenever $\overline{ab} = \bar{a}\bar{b} = \bar{0}$, then either $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$, i.e. R/P is an integral domain.

It follows in particular that a commutative ring with identity is an integral domain if and only if 0 is a prime ideal.

Corollary 14. Assume R is commutative. Every maximal ideal of R is a prime ideal.

Proof: If M is a maximal ideal then R/M is a field by Proposition 12. A field is an integral domain so the corollary follows from Proposition 13.

Examples

(1) The principal ideals generated by primes in $\mathbb{Z}$ are both prime and maximal ideals. The zero ideal in $\mathbb{Z}$ is prime but not maximal.

(2) The ideal (x) is a prime ideal in $\mathbb{Z}[x]$ since $\mathbb{Z}[x]/(x) \cong \mathbb{Z}$. This ideal is not a maximal ideal. The ideal 0 is a prime ideal in $\mathbb{Z}[x]$, but is not a maximal ideal.

EXERCISES

Let R be a ring with identity $1 \neq 0$.

1. Let L_j be the left ideal of $M_n(R)$ consisting of arbitrary entries in the jth column and zero in all other entries and let E_{ij} be the element of $M_n(R)$ whose i, j entry is 1 and whose other entries are all 0. Prove that $L_j = M_n(R)E_{ij}$ for any i. [See Exercise 6, Section 2.]
2. Assume R is commutative. Prove that the augmentation ideal in the group ring RG is generated by $\{g - 1 \mid g \in G\}$. Prove that if $G = \langle \sigma \rangle$ is cyclic then the augmentation ideal is generated by $\sigma - 1$.
3. (a) Let p be a prime and let G be an abelian group of order p^n. Prove that the nilradical of the group ring $\mathbb{F}_pG$ is the augmentation ideal (cf. Exercise 29, Section 3). [Use the preceding exercise.]

   (b) Let $G = \{g_1, \ldots, g_n\}$ be a finite group and assume R is commutative. Prove that if r is any element of the augmentation ideal of RG then $r(g_1 + \cdots + g_n) = 0$. [Use the preceding exercise.]

4. Assume R is commutative. Prove that R is a field if and only if 0 is a maximal ideal.
5. Prove that if M is an ideal such that R/M is a field then M is a maximal ideal (do not assume R is commutative).
6. Prove that R is a division ring if and only if its only left ideals are (0) and R. (The analogous result holds when "left" is replaced by "right").
7. Let R be a commutative ring with 1. Prove that the principal ideal generated by x in the polynomial ring $R[x]$ is a prime ideal if and only if R is an integral domain. Prove that (x) is a maximal ideal if and only if R is a field.
8. Let R be an integral domain. Prove that $(a) = (b)$ for some elements $a, b \in R$, if and only if $a = ub$ for some unit u of R.
9. Let R be the ring of all continuous functions on $[0, 1]$ and let I be the collection of functions $f(x)$ in R with $f(1/3) = f(1/2) = 0$. Prove that I is an ideal of R but is not a prime ideal.
10. Assume R is commutative. Prove that if P is a prime ideal of R and P contains no zero divisors then R is an integral domain.
11. Assume R is commutative. Let I and J be ideals of R and assume P is a prime ideal of R that contains $I \cap J$. Prove that either I or J is contained in P.
12. Let I be the ideal $(2, x)$ in $\mathbb{Z}[x]$. Prove that I^2 $(= II)$ contains elements which are *not* of the form ab for any $a, b \in I$ (i.e. not every element of the product IJ of two ideals can be written as a simple product ab — see Exercise 34 in Section 3 for the definition of IJ).
13. Let $\varphi : R \to S$ be a homomorphism of commutative rings.
    (a) Prove that if P is a prime ideal of S then either $\varphi^{-1}(P) = R$ or $\varphi^{-1}(P)$ is a prime ideal of R. Apply this to the special case when R is a subring of S and φ is the inclusion homomorphism to deduce that if P is a prime ideal of S then $P \cap R$ is a prime ideal of R.
    (b) Prove that if M is a maximal ideal of S and φ is surjective then $\varphi^{-1}(M)$ is a maximal ideal of R. Give an example to show that this need not be the case if φ is not surjective.
14. Assume R is commutative. Let x be an indeterminate, let $f(x)$ be a monic polynomial in $R[x]$ of degree $n \geq 1$ and use the bar notation to denote passage to the quotient ring $R[x]/(f(x))$.
    (a) Show that every element of $R[x]/(f(x))$ is of the form $\overline{p(x)}$ for some polynomial $p(x) \in R[x]$ of degree less than n, i.e.

$$R[x]/(f(x)) = \{\overline{a_0} + \overline{a_1 x} + \cdots + \overline{a_{n-1}x^{n-1}} \mid a_0, a_1, \ldots, a_{n-1} \in R\}.$$

    [If $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_0$ then $\overline{x^n} = \overline{-(b_{n-1}x^{n-1} + \cdots + b_0)}$. Use this to reduce powers of $\overline{x}$ in the quotient ring.]
    (b) Prove that if $p(x)$ and $q(x)$ are distinct polynomials in $R[x]$ which are both of degree less than n, then $\overline{p(x)} \neq \overline{q(x)}$. [Otherwise $p(x) - q(x)$ is an $R[x]$-multiple of the monic polynomial $f(x)$.]
    (c) If $f(x) = a(x)b(x)$ where both $a(x)$ and $b(x)$ have degree less than n, prove that $\overline{a(x)}$ is a zero divisor in $R[x]/(f(x))$.
    (d) If $f(x) = x^n - a$ for some nilpotent element $a \in R$, prove that $\overline{x}$ is nilpotent in $R[x]/(f(x))$.
    (e) Let p be a prime, assume $R = \mathbb{F}_p$ and $f(x) = x^p - a$ for some $a \in \mathbb{F}_p$. Prove that $\overline{x - a}$ is nilpotent in $R[x]/(f(x))$. [Use Exercise 26(c), Section 3.]
15. Let $x^2 + x + 1$ be an element of the polynomial ring $E = \mathbb{F}_2[x]$ and use the bar notation to denote passage to the quotient ring $\mathbb{F}_2[x]/(x^2 + x + 1)$.
    (a) Prove that $\overline{E}$ has 4 elements: $\overline{0}, \overline{1}, \overline{x}$ and $\overline{x+1}$.

(b) Write out the 4×4 addition table for $\overline{E}$ and deduce that the additive group $\overline{E}$ is isomorphic to the Klein 4-group.

(c) Write out the 4×4 multiplication table for $\overline{E}$ and prove that $\overline{E}^{\times}$ is isomorphic to the cyclic group of order 3. Deduce that $\overline{E}$ is a field.

16. Let $x^4 - 16$ be an element of the polynomial ring $E = \mathbb{Z}[x]$ and use the bar notation to denote passage to the quotient ring $\mathbb{Z}[x]/(x^4 - 16)$.

(a) Find a polynomial of degree ≤ 3 that is congruent to $7x^{13} - 11x^9 + 5x^5 - 2x^3 + 3$ modulo $(x^4 - 16)$.

(b) Prove that $\overline{x-2}$ and $\overline{x+2}$ are zero divisors in $\overline{E}$.

17. Let $x^3 - 2x + 1$ be an element of the polynomial ring $E = \mathbb{Z}[x]$ and use the bar notation to denote passage to the quotient ring $\mathbb{Z}[x]/(x^3 - 2x + 1)$. Let $p(x) = 2x^7 - 7x^5 + 4x^3 - 9x + 1$ and let $q(x) = (x-1)^4$.

(a) Express each of the following elements of $\overline{E}$ in the form $\overline{f(x)}$ for some polynomial $f(x)$ of degree ≤ 2: $\overline{p(x)}$, $\overline{q(x)}$, $\overline{p(x) + q(x)}$ and $\overline{p(x)q(x)}$.

(b) Prove that $\overline{E}$ is not an integral domain.

(c) Prove that $\overline{x}$ is a unit in $\overline{E}$.

18. Prove that if R is an integral domain and $R[[x]]$ is the ring of formal power series in the indeterminate x then the principal ideal generated by x is a prime ideal (cf. Exercise 3, Section 2). Prove that the principal ideal generated by x is a maximal ideal if and only if R is a field.

19. Let R be a finite commutative ring with identity. Prove that every prime ideal of R is a maximal ideal.

20. Prove that a nonzero finite commutative ring that has no zero divisors is a field (if the ring has an identity, this is Corollary 3, so do not assume the ring has a 1).

21. Prove that a finite ring with identity $1 \neq 0$ that has no zero divisors is a field (you may quote Wedderburn's Theorem).

22. Let $p \in \mathbb{Z}^+$ be a prime and let the $\mathbb{F}_p$ Quaternions be defined by

$$a + bi + cj + dk \qquad a, b, c, d \in \mathbb{Z}/p\mathbb{Z}$$

where addition is componentwise and multiplication is defined using the same relations on i, j, k as for the real Quaternions.

(a) Prove that the $\mathbb{F}_p$ Quaternions are a homomorphic image of the integral Quaternions (cf. Section 1).

(b) Prove that the $\mathbb{F}_p$ Quaternions contain zero divisors (and so they cannot be a division ring). [Use the preceding exercise.]

23. Prove that in a Boolean ring (cf. Exercise 15, Section 1) every prime ideal is a maximal ideal.

24. Prove that in a Boolean ring every finitely generated ideal is principal.

25. Assume R is commutative and for each $a \in R$ there is a positive integer n (depending on a) such that $a^n = a$. Prove that every prime ideal of R is a maximal ideal.

26. Prove that a prime ideal in a commutative ring R contains every nilpotent element (cf. Exercise 13, Section 1). Deduce that the nilradical of R (cf. Exercise 29, Section 3) is contained in the intersection of all the prime ideals of R. (Exercise 11 in Section 5 shows that the nilradical of R is equal to the intersection of all prime ideals of R).

27. Let R be a commutative ring with $1 \neq 0$. Prove that if a is a nilpotent element of R then $1 - ab$ is a unit for all $b \in R$.

28. Prove that if R is a commutative ring and $N = (a_1, a_2, \ldots, a_m)$ where each a_i is a nilpotent

element, then N is a nilpotent ideal (cf. Exercise 37, Section 3). Deduce that if the nilradical of R is finitely generated then it is a nilpotent ideal.

29. Let p be a prime and let P be a finite group of order a power of p (i.e. a p-group). Prove that the augmentation ideal in the group ring $\mathbb{Z}/p\mathbb{Z}G$ is a nilpotent ideal. (Note that this ring may be noncommutative). [Use Exercise 2.]

30. Let I be an ideal of the commutative ring R and define

$$\mathcal{R}(I) = \{r \in R \mid r^n \in I \text{ for some } n \in \mathbb{Z}^+\}$$

called the *radical* of I. Prove that $\mathcal{R}(I)$ is an ideal containing I and that $\mathcal{R}(I)/I$ is the nilradical of the quotient ring R/I, i.e. $\mathcal{R}(I)/I = \eta(R/I)$ (cf. Exercise 29, Section 3).

31. An ideal I of the commutative ring R is called a *radical ideal* if $\mathcal{R}(I) = I$.

(a) Prove that every prime ideal of R is a radical ideal.

(b) Let $n > 1$ be an integer. Prove that 0 is a radical ideal in $\mathbb{Z}/n\mathbb{Z}$ if and only if n is a product of distinct primes to the first power (i.e. n is square free). Deduce that (n) is a radical ideal of $\mathbb{Z}$ if and only if n is a product of distinct primes in $\mathbb{Z}$.

32. Let I be an ideal of the commutative ring R and define

$$\mathcal{J}(I) \text{ to be the intersection of all maximal ideals of } R \text{ that contain } I$$

where the convention is that $\mathcal{J}(R) = R$. (If I is the zero ideal, $\mathcal{J}(0)$ is called the *Jacobson radical* of the ring R, so $\mathcal{J}(I)$ is the preimage in R of the Jacobson radical of R/I.)

(a) Prove that $\mathcal{J}(I)$ is an ideal of R containing I.

(b) Prove that $\mathcal{R}(I) \subseteq \mathcal{J}(I)$, where $\mathcal{R}(I)$ is the radical of I defined in Exercise 30.

(c) Let $n > 1$ be an integer. Describe $\mathcal{J}(n\mathbb{Z})$ in terms of the prime factorization of n.

33. Let R be the ring of all continuous functions from the closed interval [0,1] to $\mathbb{R}$ and for each $c \in [0,1]$ let $M_c = \{f \in R \mid f(c) = 0\}$ (recall that M_c was shown to be a maximal ideal of R).

(a) Prove that if M is *any* maximal ideal of R then there is a real number $c \in [0,1]$ such that $M = M_c$.

(b) Prove that if b and c are distinct points in [0,1] then $M_b \neq M_c$.

(c) Prove that M_c is not equal to the principal ideal generated by $x - c$.

(d) Prove that M_c is not a finitely generated ideal.

In the preceding exercise we saw that there is a bijection[1] between the *points* of the closed interval [0,1] and the set of *maximal ideals* in the ring R of all of continuous functions on [0,1] given by $c \leftrightarrow M_c$. The next four exercises explore the generalization of this to correspondences between arbitrary subsets of [0,1] and arbitrary ideals of R. These correspondences, denoted by I : subsets of [0,1] $\to$ ideals of R and by V : ideals of $R \to$ subsets of [0,1], are not bijective (as we shall see). If, however, we restrict attention to certain subsets of [0,1] and to certain ideals in R then these correspondences become bijections which are inverses of each other. Note that in this process the underlying space [0,1] can be described as a set completely in terms of the ring theory of R (i.e. as the set of all maximal

[1] For any subset X of $\mathbb{R}$ or, more generally, for any completely regular topological space X, the map $c \mapsto M_c$ is an *injection* from X to the set of maximal ideals of R, where R is the ring of all bounded continuous real valued functions on X and M_c is the maximal ideal of functions that vanish at c. Let $\beta(X)$ be the set of maximal ideals of R. One can put a topology on $\beta(X)$ in such a way that if we identify X with its image in $\beta(X)$ then X (in its given topology) becomes a subspace of $\beta(X)$. Moreover, $\beta(X)$ is a compact space under this topology and is called the *Stone–Čech compactification* of X.

ideals of R). These exercises show that the *topology* on [0,1] (i.e. the open and closed subsets of this interval) can also be described[2] from the *ideal theory* of R. This type of correspondence, where geometric (topological) properties and problems are translated into ring theory and vice versa forms the basis for an important area of mathematics known as *algebraic geometry*. We shall discuss this process in a more abstract setting in Exercises 45 and 46 and analogous correspondences for rings of polynomials in the exercises in Section 9.5.

34. Let R be the ring of all continuous functions from the closed interval [0,1] to $\mathbb{R}$. For each subset X of [0,1] let

$$I(X) = \{f \in R \mid f(a) = 0 \text{ for all } a \in X\}$$

so $I(X)$ is the subset of R consisting of all functions that vanish on X (with the convention that $I(\emptyset) = R$).

(a) Prove that $I(X)$ is an ideal of R for all subsets X of [0,1] (hence the use of the letter I).

(b) Prove that if $X \subseteq Y$ then $I(Y) \subseteq I(X)$. Deduce that $I(X)$ is a maximal ideal of R if and only if X is a set with one element.

(c) Prove that $I(X \cup Y) = I(X) \cap I(Y)$.

35. Let R be the ring of all continuous functions from the closed interval [0,1] to $\mathbb{R}$. For each subset J of R let

$$V(J) = \{x \in [0,1] \mid f(x) = 0 \text{ for all } f \in J\}$$

so $V(J)$ is the subset of [0,1] consisting of the points where all functions in J simultaneously vanish — called the *locus* of J (with the convention that $V(\emptyset) = [0,1]$).

(a) Prove that if J is any subset of R and (J) is the ideal generated by J in R, then $V(J) = V((J))$ (so it suffices to restrict attention to ideals J in R when studying V).

(b) Prove that if $I \subseteq J$ are ideals of R then $V(J) \subseteq V(I)$.

(c) Prove that if I and J are ideals of R then $V(I + J) = V(I) \cap V(J)$.

(d) Prove that if I and J are ideals of R then $V(IJ) = V(I) \cup V(J)$ (where IJ is the ideal which is the product of I and J — cf. Exercise 34, Section 3).

36. Let R, V and I be as in the preceding two exercises.

(a) Prove that $X \subseteq V(I(X))$ for all subsets X of [0,1].

(b) Prove that $J \subseteq I(V(J))$ for all ideals J of R.

(c) Let J be an ideal of R and for each $c \in [0,1]$ let $M_c = I(\{c\})$ be the maximal ideal of functions that vanish at c. Prove that

$$\mathcal{J}(J) = \bigcap_{c \in V(J)} M_c \qquad \text{and} \qquad V(J) = \bigcup_{M_c \supseteq J} \{c\}$$

where $\mathcal{J}(J)$ is the ideal defined in Exercise 32. Deduce that $V(J) = V(\mathcal{J}(J))$. [To show the first displayed equality use Exercise 33(a).]

(d) Prove that $I(V(J)) = \mathcal{J}(J)$ for all ideals J of R. Deduce that $I(V(J)) = J$ if and only if $J = \mathcal{J}(J)$. [Use (c).]

(e) Let $X = V(J)$ for some ideal J of R. Prove that $V(I(X)) = X$. [You may use the result that for any point $c \in [0,1]$ with $c \notin X$ there is a continuous function that is 0 on X and 1 on c — this is a consequence of Urysohn's Lemma.]

[2] This is a special case of the Gelfand–Neumark Representation Theorem.

(f) A subset X of $[0,1]$ is called a *Zariski closed* set with respect to R if $X = V(J)$ for some ideal J of R. Deduce that V and I are bijections between the set of Zariski closed subsets of $[0,1]$ and the set of ideals J of R for which $J = \mathcal{J}(J)$, and show that V and I are inverse correspondences.

37. Under the notation of the preceding exercise prove that a subset X of $[0,1]$ is a Zariski closed set if and only if it is closed in the usual sense as a subset of $\mathbb{R}$ (i.e. in the usual topology on $\mathbb{R}$). [Show that a Zariski closed set is closed in the usual topology on $\mathbb{R}$ by using the fact that $f^{-1}(0)$ is a closed subset of $[0,1]$ for any continuous function f. Show the converse by using the fact that every closed subset of $[0,1]$ is a (finite or countable) intersection of closed intervals.]

38. Let $A = (a_1, a_2, \ldots, a_n)$ be a nonzero finitely generated ideal of R. Prove that there is an ideal B which is maximal with respect to the property that it does not contain A. [Use Zorn's Lemma.]

39. Assume R is commutative. Prove that the set of prime ideals in R has a minimal element with respect to inclusion (possibly the zero ideal). [Use Zorn's Lemma.]

40. A commutative ring R is called a *local ring* if it has a unique maximal ideal. Prove that if R is a local ring with maximal ideal M then every element of $R - M$ is a unit. Prove conversely that if R is a commutative ring with 1 in which the set of nonunits forms an ideal M, then R is a local ring with unique maximal ideal M.

41. Prove that the ring of all rational numbers whose denominators is odd is a local ring whose unique maximal ideal is the principal ideal generated by 2.

42. Following the notation of Exercise 26 in Section 1, let K be a field, let ν be a discrete valuation on K and let R be the valuation ring of ν. For each integer $k \geq 0$ define $A_k = \{r \in R \mid \nu(r) \geq k\} \cup \{0\}$.

(a) Prove that A_k is a principal ideal and that $A_0 \supseteq A_1 \supseteq A_2 \supseteq \cdots$.

(b) Prove that if I is any nonzero ideal of R, then $I = A_k$ for some $k \geq 0$. Deduce that R is a local ring with unique maximal ideal A_1.

43. Assume R is commutative. Prove that the following are equivalent: (see also Exercises 13 and 14 in Section 1)

(i) R has exactly one prime ideal

(ii) every element of R is either nilpotent or a unit

(iii) $R/\eta(R)$ is a field (cf. Exercise 29, Section 3).

44. A proper ideal Q of the commutative ring R is called *primary* if whenever $ab \in Q$ and $a \notin Q$ then $b^n \in Q$ for some positive integer n. (Note that the symmetry between a and b in this definition implies that if Q is a primary ideal and $ab \in Q$ with *neither* a nor b in Q, then a positive power of a and a positive power of b both lie in Q.) Establish the following facts about primary ideals.

(a) The primary ideals of $\mathbb{Z}$ are 0 and (p^n), where p is a prime and n is a positive integer.

(b) Every prime ideal of R is a primary ideal.

(c) An ideal Q of R is primary if and only if every zero divisor in R/Q is a nilpotent element of R/Q.

(d) If Q is a primary ideal then $\mathcal{R}(Q)$ is a prime ideal (cf. Exercise 30).

The next two exercises "abstract" Exercises 34 to 37 by introducing a topology on the set of prime ideals in a commutative ring R; in other words, they show how to create a topological space from a commutative ring. This topological space is called the *prime spectrum* of R and the topology on the spectrum is called the *Zariski topology*.

45. Let R be commutative and let $\mathrm{Spec}(R)$ be the set of all prime ideals of R. For any subset X of R let $V(X)$ be the set of all prime ideals of R that contain X.

(a) Prove that $V(X) = V((X))$ for any subset X of R (where (X) is, as usual, the ideal of R generated by X).

(b) Prove that $V(0) = \mathrm{Spec}(R)$ and $V(R) = \emptyset$.

(c) Prove that if X_i, $i \in I$ is any collection of subsets of R, then $V(\cup_{i\in I} X_i) = \cap_{i\in I} V(X_i)$.

(d) Prove that if X and Y are any two of subsets of R, then $V(X \cap Y) = V(X) \cup V(Y)$.

Parts (b) to (d) show that the sets $V(X)$, as X runs over all subsets of R, satisfy the axioms for a collection of closed sets in a topological space. The subsets $V(X)$ of $\mathrm{Spec}(R)$ are called *Zariski closed* sets. Henceforth, $\mathrm{Spec}(R)$ is considered to have the topology defined by taking the Zariski closed sets to be the closed sets — this is the Zariski topology on $\mathrm{Spec}(R)$.

46. Let R and S be commutative rings and let $\varphi : R \to S$ be a ring homomorphism such that $\varphi(1) = 1$. By Exercise 13, if P is a prime ideal of S, then $\varphi^{-1}(P)$ is a prime ideal of R. Thus φ induces a map

$$\varphi^* : \mathrm{Spec}(S) \to \mathrm{Spec}(R) \qquad \text{defined by} \qquad \varphi^*(P) = \varphi^{-1}(P) \text{ for all } P \in \mathrm{Spec}(S).$$

(a) Prove that for any ideal J of R, $\varphi^{*-1}(V(J)) = V((\varphi(J)))$ (where $(\varphi(J))$ is the ideal generated by $\varphi(J)$ in S and V is as defined in the preceding exercise). Deduce that φ^* is a *continuous map* with respect to the Zariski topologies on $\mathrm{Spec}(S)$ and $\mathrm{Spec}(R)$. [A map is continuous if the inverse image of a closed set is closed.]

(b) If $\psi : S \to T$ is also a homomorphism of commutative rings, prove that $(\psi \circ \varphi)^* = \varphi^* \circ \psi^*$.

7.5 RINGS OF FRACTIONS

Throughout this section R is a commutative ring. Proposition 2 shows that if a is not zero nor a zero divisor and $ab = ac$ in R then $b = c$. Thus a nonzero element that is not a zero divisor enjoys some of the properties of a unit without necessarily possessing a multiplicative inverse in R. On the other hand, we saw in Section 1 that a zero divisor a cannot be a unit in R and, by definition, if a is a zero divisor we cannot always cancel the a's in the equation $ab = ac$ to obtain $b = c$ (take $c = 0$ for example). The aim of this section is to prove that a commutative ring R is always a subring of a larger ring Q in which every nonzero element of R that is not a zero divisor is a unit in Q. The principal application of this will be to integral domains, in which case this ring Q will be a field — called its *field of fractions* or *quotient field*. Indeed, the paradigm for the construction of Q from R is the one offered by the construction of the field of rational numbers from the integral domain $\mathbb{Z}$.

In order to see the essential features of the construction of the field $\mathbb{Q}$ from the integral domain $\mathbb{Z}$ we review the basic properties of fractions. Each rational number may be represented in many different ways as the quotient of two integers (for example, $\frac{1}{2} = \frac{2}{4} = \frac{3}{6} = \ldots$, etc.). These representations are related by

$$\frac{a}{b} = \frac{c}{d} \qquad \text{if and only if} \qquad ad = bc.$$

In more precise terms, the fraction $\frac{a}{b}$ is the equivalence class of ordered pairs (a, b) of integers with $b \neq 0$ under the equivalence relation: $(a, b) \sim (c, d)$ if and only if $ad = bc$. The arithmetic operations on fractions are given by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad \text{and} \qquad \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}.$$

These are well defined (independent of choice of representatives of the equivalence classes) and make the set of fractions into a commutative ring (in fact, a field), $\mathbb{Q}$. The integers $\mathbb{Z}$ are identified with the subring $\{\frac{a}{1} \mid a \in \mathbb{Z}\}$ of $\mathbb{Q}$ and every nonzero integer a has an inverse $\frac{1}{a}$ in $\mathbb{Q}$.

It seems reasonable to attempt to follow the same steps for any commutative ring R, allowing arbitrary denominators. If, however, b is zero or a zero divisor in R, say $bd = 0$, and if we allow b as a denominator, then we should expect to have

$$d = \frac{d}{1} = \frac{bd}{b} = \frac{0}{b} = 0$$

in the "ring of fractions" (where, for convenience, we have assumed R has a 1). Thus if we allow zero or zero divisors as denominators there must be some collapsing in the sense that we cannot expect R to appear naturally as a subring of this "ring of fractions." A second restriction is more obviously imposed by the laws of addition and multiplication: if ring elements b and d are allowed as denominators, then bd must also be a denominator, i.e. the set of denominators must be closed under multiplication in R. The main result of this section shows that these two restrictions are sufficient to construct a ring of fractions for R. Note that this theorem includes the construction of $\mathbb{Q}$ from $\mathbb{Z}$ as a special case.

Theorem 15. Let R be a commutative ring. Let D be any nonempty subset of R that does not contain 0, does not contain any zero divisors and is closed under multiplication (i.e. $ab \in D$ for all $a, b \in D$). Then there is a commutative ring Q with 1 such that Q contains R as a subring and every element of D is a unit in Q. The ring Q has the following additional properties.

(1) every element of Q is of the form rd^{-1} for some $r \in R$ and $d \in D$. In particular, if $D = R - \{0\}$ then Q is a field.

(2) (uniqueness of Q) The ring Q is the *"smallest"* ring containing R in which all elements of D become units, in the following sense. Let S be any commutative ring with identity and let $\varphi : R \to S$ be any injective ring homomorphism such that $\varphi(d)$ is a unit in S for every $d \in D$. Then there is an injective homomorphism $\Phi : Q \to S$ such that $\Phi|_R = \varphi$. In other words, any ring containing an isomorphic copy of R in which all the elements of D become units must also contain an isomorphic copy of Q.

Remark: In Exercise 6 a more general construction is outlined. Readers wishing to achieve the greatest degree of generality at the outset may read this exercise and follow the steps of the proof below under its weaker hypotheses.

Proof: Let $\mathcal{F} = \{(r,d) \mid r \in R,\ d \in D\}$ and define the relation $\sim$ on $\mathcal{F}$ by

$$(r,d) \sim (s,e) \qquad \text{if and only if} \qquad re = sd.$$

It is immediate that this relation is reflexive and symmetric. Suppose $(r,d) \sim (s,e)$ and $(s,e) \sim (t,f)$. Then $re - sd = 0$ and $sf - te = 0$. Multiplying the first of these equations by f and the second by d and adding them gives: $(rf - td)e = 0$. Since $e \in D$ is neither zero nor a zero divisor we must have $rf - td = 0$, i.e. $(r,d) \sim (t,f)$. This proves $\sim$ is transitive, hence an equivalence relation. Denote the equivalence class of (r,d) by $\dfrac{r}{d}$:

$$\frac{r}{d} = \{(a,b) \mid a \in R,\ b \in D \text{ and } rb = ad\}.$$

Let Q be the set of equivalence classes under $\sim$. Note that $\dfrac{r}{d} = \dfrac{re}{de}$ in Q for all $e \in D$, since D is closed under multiplication.

We now define an additive and multiplicative structure on Q:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd} \qquad \text{and} \qquad \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}.$$

In order to prove that Q is a commutative ring with identity there are a number of things to check:

(1) these operations are well defined (i.e. do not depend on the choice of representatives for the equivalence classes),

(2) Q is an abelian group under addition, where the additive identity is $\dfrac{0}{d}$ for any $d \in D$ and the additive inverse of $\dfrac{a}{d}$ is $\dfrac{-a}{d}$,

(3) multiplication is associative, distributive and commutative, and

(4) Q has an identity ($= \dfrac{d}{d}$ for any $d \in D$).

These are all completely straightforward calculations involving only arithmetic in R and the definition of $\sim$. Again we need that D is closed under multiplication for addition and multiplication to be defined.

For example, to check that addition is well defined assume $\dfrac{a}{b} = \dfrac{a'}{b'}$ (i.e. $ab' = a'b$) and $\dfrac{c}{d} = \dfrac{c'}{d'}$ (i.e. $cd' = c'd$). We must show that $\dfrac{ad+bc}{bd} = \dfrac{a'd'+b'c'}{b'd'}$, i.e.

$$(ad+bc)(b'd') = (a'd'+b'c')(bd).$$

The left hand side of this equation is $ab'dd' + cd'bb'$ substituting $a'b$ for ab' and $c'd$ for cd' gives $a'bdd' + c'dbb'$, which is the right hand side. Hence addition of fractions is well defined. Checking the details in the other parts of (1) to (4) involves even easier manipulations and so is left as an exercise.

Next we embed R into Q by defining

$$\iota : R \to Q \qquad \text{by} \qquad \iota : r \mapsto \frac{rd}{d} \quad \text{where } d \text{ is any element of } D.$$

Since $\frac{rd}{d} = \frac{re}{e}$ for all $d, e \in D$, $\iota(r)$ does not depend on the choice of $d \in D$. Since D is closed under multiplication, one checks directly that ι is a ring homomorphism. Furthermore, ι is injective because

$$\iota(r) = 0 \Leftrightarrow \frac{rd}{d} = \frac{0}{d} \Leftrightarrow rd^2 = 0 \Leftrightarrow r = 0$$

because d (hence also d^2) is neither zero nor a zero divisor. The subring $\iota(R)$ of Q is therefore isomorphic to R. We henceforth identify each $r \in R$ with $\iota(r)$ and so consider R as a subring of Q.

Next note that each $d \in D$ has a multiplicative inverse in Q: namely, if d is represented by the fraction $\frac{de}{e}$ then its multiplicative inverse is $\frac{e}{de}$. One then sees that every element of Q may be written as $r \cdot d^{-1}$ for some $r \in R$ and some $d \in D$. In particular, if $D = R - \{0\}$, every nonzero element of Q has a multiplicative inverse and Q is a field.

It remains to establish the uniqueness property of Q. Assume $\varphi : R \to S$ is an injective ring homomorphism such that $\varphi(d)$ is a unit in S for all $d \in D$. Extend φ to a map $\Phi : Q \to S$ by defining $\Phi(rd^{-1}) = \varphi(r)\varphi(d)^{-1}$ for all $r \in R$, $d \in D$. This map is well defined, since $rd^{-1} = se^{-1}$ implies $re = sd$, so $\varphi(r)\varphi(e) = \varphi(s)\varphi(d)$, and then

$$\Phi(rd^{-1}) = \varphi(r)\varphi(d)^{-1} = \varphi(s)\varphi(e)^{-1} = \Phi(se^{-1}).$$

It is straightforward to check that Φ is a ring homomorphism — the details are left as an exercise. Finally, Φ is injective because $rd^{-1} \in \ker \Phi$ implies $r \in \ker \Phi \cap R = \ker \varphi$; since φ is injective this forces r and hence also rd^{-1} to be zero. This completes the proof.

Definition. Let R, D and Q be as in Theorem 15.

(1) The ring Q is called the *ring of fractions* of D with respect to R and is denoted $D^{-1}R$.

(2) If R is an integral domain and $D = R - \{0\}$, Q is called the *field of fractions* or *quotient field* of R.

If A is a subset of a field F (for example, if A is a subring of F), then the intersection of all the subfields of F containing A is a subfield of F and is called the subfield *generated* by A. This subfield is the smallest subfield of F containing A (namely, any subfield of F containing A contains the subfield generated by A).

The next corollary shows that the smallest field containing an integral domain R is its field of fractions.

Corollary 16. Let R be an integral domain and let Q be the field of fractions of R. If a field F contains a subring R' isomorphic to R then the subfield of F generated by R' is isomorphic to Q.

Proof: Let $\varphi : R \cong R' \subseteq F$ be a (ring) isomorphism of R to R'. In particular, $\varphi : R \to F$ is an injective homomorphism from R into the field F. Let $\Phi : Q \to F$ be the extension of φ to Q as in the theorem. By Theorem 15, Φ is injective, so $\Phi(Q)$ is an isomorphic copy of Q in F containing $\varphi(R) = R'$. Now, any subfield of F containing $R' = \varphi(R)$ contains the elements $\varphi(r_1)\varphi(r_2)^{-1} = \varphi(r_1 r_2^{-1})$ for all $r_1, r_2 \in R$. Since every element of Q is of the form $r_1 r_2^{-1}$ for some $r_1, r_2 \in R$, it follows that any subfield of F containing R' contains the field $\Phi(Q)$, so that $\Phi(Q)$ is the subfield of F generated by R', proving the corollary.

Examples

(1) If R is a field then its field of fractions is just R itself.

(2) The integers $\mathbb{Z}$ are an integral domain whose field of fractions is the field $\mathbb{Q}$ of rational numbers.

(3) The subring $2\mathbb{Z}$ of $\mathbb{Z}$ also has no zero divisors (but has no identity). Its field of fractions is also $\mathbb{Q}$. Note how an identity "appears" in the field of fractions.

(4) If R is any integral domain, then the polynomial ring $R[x]$ is also an integral domain. The associated field of fractions is the field of *rational functions* in the variable x over R. The elements of this field are of the form $\dfrac{p(x)}{q(x)}$, where $p(x)$ and $q(x)$ are polynomials with coefficients in R with $q(x)$ not the zero polynomial. In particular, $p(x)$ and $q(x)$ may both be constant polynomials, so the field of rational functions contains the field of fractions of R: elements of the form $\dfrac{a}{b}$ such that $a, b \in R$ and $b \neq 0$. If F is a field, we shall denote the field of rational functions by $F(x)$. Thus if F is the field of fractions of the integral domain R then the field of rational functions over R is the same as the field of rational functions over F, namely $F(x)$.

For example, suppose $R = \mathbb{Z}$, so $F = \mathbb{Q}$. If $p(x), q(x)$ are polynomials in $\mathbb{Q}[x]$ then for some integer N, $Np(x), Nq(x)$ have integer coefficients (let N be a common denominator for all the coefficients in $p(x)$ and $q(x)$, for example). Then $\dfrac{p(x)}{q(x)} = \dfrac{Np(x)}{Nq(x)}$ can be written as the quotient of two polynomials with integer coefficients, so the field of fractions of $\mathbb{Q}[x]$ is the same as the field of fractions of $\mathbb{Z}[x]$.

(5) If R is any commutative ring with identity and d is neither zero nor a zero divisor in R we may form the ring $R[\frac{1}{d}]$ by setting $D = \{1, d, d^2, d^3, \dots\}$ and defining $R[\frac{1}{d}]$ to be the ring of fractions $D^{-1}R$. Note that R is the subring of elements of the form $\frac{r}{1}$. In this way any nonzero element of R that is not a zero divisor can be inverted in a larger ring containing R. Note that the elements of $R[\frac{1}{d}]$ look like polynomials in $\frac{1}{d}$ with coefficients in R, which explains the notation.

EXERCISES

Let R be a commutative ring with identity $1 \neq 0$.

1. Fill in all the details in the proof of Theorem 15.

2. Let R be an integral domain and let D be a nonempty subset of R that is closed under multiplication. Prove that the ring of fractions $D^{-1}R$ is isomorphic to a subring of the quotient field of R (hence is also an integral domain).

3. Let F be a field. Prove that F contains a unique smallest subfield F_0 and that F_0 is isomor-

phic to either $\mathbb{Q}$ or $\mathbb{Z}/p\mathbb{Z}$ for some prime p (F_0 is called the *prime subfield* of F). [See Exercise 26, Section 3.]

4. Prove that any subfield of $\mathbb{R}$ must contain $\mathbb{Q}$.
5. Let R be an integral domain with field of fractions F and let $R[[x]]$ be the ring of formal power series in the indeterminate x with coefficients from R (cf. Exercise 3 in Section 2). Prove that the field of fractions of $R[[x]]$ is the ring $F((x))$ of formal Laurent series as defined in Exercise 5 of Section 2.

The following exercise outlines the general construction for rings of fractions (this process is also called *localization* of the ring R). Some of the subsequent exercises utilize this construction.

6. Let D be any nonempty subset of R that is closed under multiplication. For convenience assume $1 \in D$.
   (a) Define a relation on $\mathcal{F} = \{(r, d) \mid r \in R,\ d \in D\}$ by
   $$(r, d) \sim (s, e) \quad \text{if and only if} \quad (re - sd)f = 0 \text{ for some } f \in D.$$
   Prove that $\sim$ is an equivalence relation.
   (b) Denote the equivalence class of $(r, d) \in \mathcal{F}$ by $\frac{r}{d}$ and let $D^{-1}R$ be the set of equivalence classes. Define addition and multiplication of equivalence classes by the same rules as in Theorem 15. Prove that $D^{-1}R$ is a commutative ring with 1.
   (c) Prove that the map $\iota : R \to D^{-1}R$ defined by $\iota(r) = \frac{r}{1}$ is a ring homomorphism and $\ker \iota = \{r \in R \mid rd = 0 \text{ for some } d \in D\}$.
   (d) Prove that $D^{-1}R$ is the zero ring if and only if $0 \in D$.
   (e) Establish the following uniqueness property of $D^{-1}R$. Let $\varphi : R \to S$ be any homomorphism of R into a commutative ring S with identity such that $\varphi(1) = 1$ and $\varphi(d)$ is a unit in S for all $d \in D$. Show that there is a unique ring homomorphism $\psi : D^{-1}R \to S$ such that $\varphi = \psi \circ \iota$.
7. Let $R = \mathbb{Z}/6\mathbb{Z}$ and let $D = \{1, 2, 4\}$. In the notation of the preceding exercise, for this particular choice of R and D:
   (a) write out all the elements of $\mathcal{F}$,
   (b) find the equivalence classes under $\sim$ (i.e. write out the elements of $D^{-1}R$),
   (c) write out $\iota(r)$ for all elements $r \in R$,
   (d) find the kernel of ι, and
   (e) prove that $D^{-1}R$ is isomorphic to $\mathbb{Z}/3\mathbb{Z}$ (so $D^{-1}R$ is an integral domain even though R is not an integral domain).
8. Let D be a nonempty subset of R containing 1 that is closed under multiplication and let Q be the ring of fractions $D^{-1}R$ (if one wishes to avoid using Exercise 6, assume also that D does not contain 0 nor any zero divisor). Let ι be the map from R to Q defined in Theorem 15 or in Exercise 6.
   (a) Prove that the correspondence
   $$I \text{ an ideal of } R \quad \longrightarrow \quad \text{the ideal in } Q \text{ generated by } \iota(I)$$
   from the ideals in R to the ideals in Q is a surjection (so every ideal of $D^{-1}R$ "comes from" an ideal in R).
   (b) Use the example $R = \mathbb{Z}$, $D = \mathbb{Z} - \{0\}$ to show that the correspondence
   $$\iota^{-1}(I') \quad \longleftarrow \quad I'$$
   from the ideals in Q to the ideals in R need not be surjective.

(c) Prove that the correspondences in (a) and (b) do give a bijection between the set of *prime ideals* of R that are disjoint from D and the set of *prime ideals* of Q ((b) shows this need not be the case if the qualifier "prime" is removed).

9. Let P be a prime ideal of R and let D be the set $R - P$ (note that D is closed under multiplication by definition of a prime ideal). Prove that the ring of fractions $D^{-1}R$ has a unique maximal ideal i.e. is a local ring — cf. Exercise 40, Section 4. (The ring $D^{-1}R$ is called the *localization of R at P* and is denoted by R_P).

10. Let p be a prime in $\mathbb{Z}$ and let $\mathbb{Z}_{(p)}$ denote the localization of $\mathbb{Z}$ at the prime ideal $(p) = p\mathbb{Z}$ as in the previous exercise.

(a) Show that $\mathbb{Z}_{(p)}$ is (isomorphic to) the subring of $\mathbb{Q}$ consisting of the rational numbers whose denominators (in lowest terms) are not divisible by p.

(b) Prove that the proper ideals of $\mathbb{Z}_{(p)}$ are (0) and the ideals (p), (p^2), $(p^3)\ldots$. [Use Exercise 8.]

(c) Prove that $\mathbb{Z}_{(p)}$ is not isomorphic to $\mathbb{Z}_{(q)}$ for distinct primes p and q.

11. Recall from Exercise 29 of Section 3 that the nilradical of R is the set of all nilpotent elements of R. Prove that the nilradical of R is the intersection of all prime ideals of R. [Recall from Exercise 26 of Section 4 that the nilradical is contained in the intersection of all prime ideals. Show conversely that if d is an element of R which is not nilpotent then there is a prime ideal of R not containing d as follows. Let $D = \{1, d, d^2, \ldots\}$, let P be a maximal ideal of $D^{-1}R$ and apply Exercise 13(a) of Section 4 to $\iota^{-1}(P)$. Alternatively, one may avoid using localization by showing directly that an ideal of R which is maximal with respect to not containing d exists (by Zorn's Lemma) and is prime.]

12. Prove that if I is an ideal of R then the radical of I, $\mathcal{R}(I)$, is the intersection of all prime ideals of R that contain I (cf. Exercise 30 of Section 4). [Use the preceding exercise.]

13. Prove that if I is a primary ideal of R (cf. Exercise 44 of Section 4) then the radical of I is the smallest (in the sense of inclusion) prime ideal containing I. [Use Exercise 44 of Section 4 and the preceding exercise.]

14. Prove that the real numbers, $\mathbb{R}$, contains a subring A with $1 \in A$ and A maximal (under inclusion) with respect to the property that $\frac{1}{2} \notin A$. [Use Zorn's Lemma.] (It can be shown that $\mathbb{R}$ is the quotient field of A, so $\mathbb{R}$ is the quotient field of some proper subring.)

7.6 THE CHINESE REMAINDER THEOREM

Throughout this section we shall assume unless otherwise stated that all rings are commutative with an identity $1 \neq 0$.

Given an arbitrary collection of rings (not necessarily satisfying the conventions above), their *(ring) direct product* is defined to be their direct product as (abelian) groups made into a ring by defining multiplication componentwise. In particular, if R_1 and R_2 are two rings, we shall denote by $R_1 \times R_2$ their direct product (as rings), that is, the set of ordered pairs (r_1, r_2) with $r_1 \in R_1$ and $r_2 \in R_2$ where addition and multiplication are performed componentwise:

$$(r_1, r_2) + (s_1, s_2) = (r_1 + s_1, r_2 + s_2) \qquad \text{and} \qquad (r_1, r_2)(s_1, s_2) = (r_1 s_1, r_2 s_2).$$

We note that a map φ from a ring R into a direct product ring is a homomorphism if and only if the induced maps into each of the components is a homomorphism.

There is a generalization to arbitrary rings of the notion in $\mathbb{Z}$ of two integers n and m being relatively prime (even to rings where the notion of greatest common

divisor is not defined). In $\mathbb{Z}$ this is equivalent to being able to solve the equation $nx + my = 1$ in integers x and y (this fact was stated in Chapter 0 and will be proved in Chapter 8). This in turn is equivalent to $n\mathbb{Z} + m\mathbb{Z} = \mathbb{Z}$ as ideals (in general, $n\mathbb{Z} + m\mathbb{Z} = (m, n)\mathbb{Z}$). This motivates the definition:

Definition. The ideals A and B of the ring R are said to be *comaximal* if $A + B = R$.

Recall that the *product*, AB, of the ideals A and B of R is the ideal consisting of all finite sums of elements of the form xy, $x \in A$ and $y \in B$ (cf. Exercise 34, Section 3). If $A = (a)$ and $B = (b)$, then $AB = (ab)$. More generally, the product of the ideals $A_1, A_2, \ldots, A_k$ is the ideal of all finite sums of elements of the form $x_1 x_2 \cdots x_k$ such that $x_i \in A_i$ for all i. If $A_i = (a_i)$, then $A_1 \cdots A_k = (a_1 \cdots a_k)$.

Theorem 17. (The Chinese Remainder Theorem) Let $A_1, A_2, \ldots, A_k$ be ideals in R. The map

$$R \to R/A_1 \times R/A_2 \times \cdots \times R/A_k \qquad \text{defined by} \qquad r \mapsto (r + A_1, r + A_2, \ldots, r + A_k)$$

is a ring homomorphism with kernel $A_1 \cap A_2 \cap \cdots \cap A_k$. If for each $i, j \in \{1, 2, \ldots, k\}$ with $i \neq j$ the ideals A_i and A_j are comaximal, then this map is surjective and $A_1 \cap A_2 \cap \cdots \cap A_k = A_1 A_2 \cdots A_k$, so

$$R/(A_1 A_2 \cdots A_k) = R/(A_1 \cap A_2 \cap \cdots \cap A_k) \cong R/A_1 \times R/A_2 \times \cdots \times R/A_k.$$

Proof: We first prove this for $k = 2$; the general case will follow by induction. Let $A = A_1$ and $B = A_2$. Consider the map $\varphi : R \to R/A \times R/B$ defined by $\varphi(r) = (r \bmod A, r \bmod B)$, where mod A means the class in R/A containing r (that is, $r + A$). This map is a ring homomorphism because φ is just the natural projection of R into R/A and R/B for the two components. The kernel of φ consists of all the elements $r \in R$ that are in A and in B, i.e. $A \cap B$. To complete the proof in this case it remains to show that when A and B are comaximal, φ is surjective and $A \cap B = AB$. Since $A + B = R$, there are elements $x \in A$ and $y \in B$ such that $x + y = 1$. This equation shows that $\varphi(x) = (0, 1)$ and $\varphi(y) = (1, 0)$ since, for example, x is an element of A and $x = 1 - y \in 1 + B$. If now $(r_1 \bmod A, r_2 \bmod B)$ is an arbitrary element in $R/A \times R/B$, then the element $r_2 x + r_1 y$ maps to this element since

$$\begin{aligned}
\varphi(r_2 x + r_1 y) &= \varphi(r_2)\varphi(x) + \varphi(r_1)\varphi(y) \\
&= (r_2 \bmod A, r_2 \bmod B)(0, 1) + (r_1 \bmod A, r_1 \bmod B)(1, 0) \\
&= (0, r_2 \bmod B) + (r_1 \bmod A, 0) \\
&= (r_1 \bmod A, r_2 \bmod B).
\end{aligned}$$

This shows that φ is indeed surjective. Finally, the ideal AB is always contained in $A \cap B$. If A and B are comaximal and x and y are as above, then for any $c \in A \cap B$, $c = c1 = cx + cy \in AB$. This establishes the reverse inclusion $A \cap B \subseteq AB$ and completes the proof when $k = 2$.

The general case follows easily by induction from the case of two ideals using $A = A_1$ and $B = A_2 \cdots A_k$ once we show that A_1 and $A_2 \cdots A_k$ are comaximal. By

hypothesis, for each $i \in \{2, 3, \ldots, k\}$ there are elements $x_i \in A_1$ and $y_i \in A_i$ such that $x_i + y_i = 1$. Since $x_i + y_i \equiv y_i \bmod A_1$, it follows that $1 = (x_2 + y_2) \cdots (x_k + y_k) \in A_1 + (A_2 \cdots A_k)$. This completes the proof.

This theorem obtained its name from the special case $\mathbb{Z}/mn\mathbb{Z} \cong (\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ *as rings* when m and n are relatively prime integers. We proved this isomorphism just for the additive groups earlier. This isomorphism, phrased in number-theoretic terms, relates to simultaneously solving two congruences modulo relatively prime integers (and states that such congruences can always be solved, and uniquely). Such problems were considered by the ancient Chinese, hence the name. Some examples are provided in the exercises.

Since the isomorphism in the Chinese Remainder Theorem is an isomorphism of *rings*, in particular the groups of *units* on both sides must be isomorphic. It is easy to see that the units in any direct product of rings are the elements that have units in each of the coordinates. In the case of $\mathbb{Z}/mn\mathbb{Z}$ the Chinese Remainder Theorem gives the following isomorphism on the groups of units:

$$(\mathbb{Z}/mn\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times.$$

More generally we have the following result.

Corollary 18. Let n be a positive integer and let ${p_1}^{\alpha_1}{p_2}^{\alpha_2}\ldots{p_k}^{\alpha_k}$ be its factorization into powers of distinct primes. Then

$$\mathbb{Z}/n\mathbb{Z} \cong (\mathbb{Z}/{p_1}^{\alpha_1}\mathbb{Z}) \times (\mathbb{Z}/{p_2}^{\alpha_2}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/{p_k}^{\alpha_k}\mathbb{Z})$$

as rings, so in particular we have the following isomorphism of multiplicative groups:

$$(\mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/{p_1}^{\alpha_1}\mathbb{Z})^\times \times (\mathbb{Z}/{p_2}^{\alpha_2}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/{p_k}^{\alpha_k}\mathbb{Z})^\times.$$

If we compare orders on the two sides of this last isomorphism, we obtain the formula

$$\varphi(n) = \varphi({p_1}^{\alpha_1})\varphi({p_2}^{\alpha_2})\ldots\varphi({p_k}^{\alpha_k})$$

for the Euler φ–function. This in turn implies that φ is what in elementary number theory is termed a *multiplicative function*, namely that $\varphi(ab) = \varphi(a)\varphi(b)$ whenever a and b are relatively prime positive integers. The value of φ on prime powers p^α is easily seen to be $\varphi(p^\alpha) = p^{\alpha-1}(p-1)$ (cf. Chapter 0). From this and the multiplicativity of φ we obtain its value on all positive integers.

Corollary 18 is also a step toward a determination of the decomposition of the abelian group $(\mathbb{Z}/n\mathbb{Z})^\times$ into a direct product of cyclic groups. The complete structure is derived at the end of Section 9.5.

EXERCISES

Let R be a ring with identity $1 \neq 0$.

1. An element $e \in R$ is called an *idempotent* if $e^2 = e$. Assume e is an idempotent in R and $er = re$ for all $r \in R$. Prove that Re and $R(1-e)$ are two-sided ideals of R and that $R \cong Re \times R(1-e)$. Show that e and $1-e$ are identities for the subrings Re and $R(1-e)$ respectively.

2. Let R be a finite Boolean ring with identity $1 \neq 0$ (cf. Exercise 15 of Section 1). Prove that $R \cong \mathbb{Z}/2\mathbb{Z} \times \cdots \times \mathbb{Z}/2\mathbb{Z}$. [Use the preceding exercise.]
3. Let R and S be rings with identities. Prove that every ideal of $R \times S$ is of the form $I \times J$ where I is an ideal of R and J is an ideal of S.
4. Prove that if R and S are nonzero rings then $R \times S$ is never a field.
5. Let $n_1, n_2, \ldots, n_k$ be integers which are relatively prime in pairs (i.e. $(n_i, n_j) = 1$ for all $i \neq j$).
   (a) Show that the Chinese Remainder Theorem implies that for any $a_1, \ldots, a_k \in \mathbb{Z}$ there is a solution $x \in \mathbb{Z}$ to the simultaneous congruences
   $$\begin{aligned} x &\equiv a_1 \bmod n_1 \\ x &\equiv a_2 \bmod n_2 \\ &\vdots \\ x &\equiv a_k \bmod n_k \end{aligned}$$
   and that the solution x is unique mod $n = n_1 n_2 \ldots n_k$.
   (b) Let $n_i' = n/n_i$ be the quotient of n by n_i, which is relatively prime to n_i by assumption. Let t_i be the inverse of n_i' mod n_i. Prove that the solution x in (a) is given by
   $$x = a_1 t_1 n_1' + a_2 t_2 n_2' + \cdots + a_k t_k n_k' \bmod n.$$
   Note that the elements t_i can be quickly found by the Euclidean Algorithm as described in Section 0.2 (writing $an_i + bn_i' = (n_i, n_i') = 1$ gives $t_i = b$) and that these then quickly give the solutions to the system of congruences above for any choice of $a_1, a_2, \ldots, a_k$.
   (c) Solve the system of congruences
   $$\begin{aligned} x &\equiv 1 \bmod 8 \\ x &\equiv 2 \bmod 25 \\ x &\equiv 3 \bmod 81 \end{aligned}$$
   and the system
   $$\begin{aligned} y &\equiv 5 \bmod 8 \\ y &\equiv 12 \bmod 25 \\ y &\equiv 47 \bmod 81\,. \end{aligned}$$
6. Let $f_1(x), f_2(x), \ldots, f_k(x)$ be polynomials with integer coefficients of the same degree d. Let $n_1, n_2, \ldots, n_k$ be integers which are relatively prime in pairs (i.e. $(n_i, n_j) = 1$ for all $i \neq j$). Use the Chinese Remainder Theorem to prove there exists a polynomial $f(x)$ with integer coefficients and of degree d with
   $$\begin{aligned} f(x) &\equiv f_1(x) \bmod n_1 \\ f(x) &\equiv f_2(x) \bmod n_2 \\ &\vdots \\ f(x) &\equiv f_k(x) \bmod n_k \end{aligned}$$
   i.e. the coefficients of $f(x)$ agree with the coefficients of $f_i(x)$ mod n_i. Show that if all the $f_i(x)$ are monic, then $f(x)$ may also be chosen monic. [Apply the Chinese Remainder Theorem in $\mathbb{Z}$ to each of the coefficients separately.]
7. Let m and n be positive integers with n dividing m. Prove that the natural surjective ring projection $\mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is also surjective on the units: $(\mathbb{Z}/m\mathbb{Z})^\times \to (\mathbb{Z}/n\mathbb{Z})^\times$.

8

Euclidean Domains, Principal Ideal Domains, and Unique Factorization Domains

There are a number of classes of rings with more algebraic structure than generic rings. Those considered in this chapter are rings with a division algorithm (Euclidean Domains), rings in which every ideal is principal (Principal Ideal Domains) and rings in which elements have factorizations into primes (Unique Factorization Domains). The principal examples of such rings are the ring $\mathbb{Z}$ of integers and polynomial rings $F[x]$ with coefficients in some field F. We prove here all the theorems on the integers $\mathbb{Z}$ stated in Chapter 0 as special cases of results valid for more general rings. These results will be applied to the special case of the ring $F[x]$ in the next chapter.

All rings in this chapter are commutative.

8.1 EUCLIDEAN DOMAINS

We first define the notion of a *norm* on an integral domain R. This is essentially no more than a measure of "size" in R.

Definition. Any function $N : R \to \mathbb{Z}^+ \cup \{0\}$ with $N(0) = 0$ is called a *norm* on the integral domain R.

We observe that this notion of a norm is fairly weak and that it is possible for the same integral domain R to possess several different norms.

Definition. The integral domain R is said to be a *Euclidean Domain* (or possess a *Division Algorithm*) if there is a norm N on R such that for any two elements a and b of R with $b \neq 0$ there exist elements q and r in R with

$$a = qb + r \qquad \text{with } r = 0 \text{ or } N(r) < N(b).$$

The element q is called the *quotient* and the element r the *remainder* of the division.

The importance of the existence of a Division Algorithm on an integral domain R is that it allows a *Euclidean Algorithm* for two elements a and b of R: by successive "divisions" (these actually *are* divisions in the field of fractions of R) we can write

$$a = q_0 b + r_0 \tag{0}$$

$$b = q_1 r_0 + r_1 \tag{1}$$

$$r_0 = q_2 r_1 + r_2 \tag{2}$$

$$\vdots$$

$$r_{n-2} = q_n r_{n-1} + r_n \tag{n}$$

$$r_{n-1} = q_{n+1} r_n \tag{n+1}$$

where r_n is the last nonzero remainder. Such an r_n exists since $N(b) > N(r_0) > N(r_1) > \cdots > N(r_n)$ is a decreasing sequence of nonnegative integers if the remainders are nonzero, and such a sequence cannot continue indefinitely. Note also that there is no guarantee that these elements are *unique*.

Examples

(0) Fields are trivial examples of Euclidean Domains where any norm will satisfy the defining condition (e.g. $N(a) = 0$ for all a). This is because for every a, b with $b \neq 0$ we have $a = qb + 0$, where $q = ab^{-1}$.

(1) The integers $\mathbb{Z}$ are a Euclidean Domain with norm given by $N(a) = |a|$, the usual absolute value. The existence of a Division Algorithm in $\mathbb{Z}$ (the familiar "long division" of elementary arithmetic) is verified as follows. Let a and b be two nonzero integers and suppose first that $b > 0$. The half open intervals $[nb, (n+1)b)$, $n \in \mathbb{Z}$ partition the real line and so a is in one of them: say $a \in [kb, (k+1)b)$. For $q = k$ we have $a - qb = r \in [0, |b|)$ as needed. If $b < 0$ (so $-b > 0$), by what we have just seen there is an integer q such that $a = q(-b) + r$ with either $r = 0$ or $|r| < |-b|$; then $a = (-q)b + r$ satisfies the requirements of the Division Algorithm for a and b. This argument can be made more formal by using induction on $|a|$.

Note that if a is not a multiple of b there are always two possibilities for the pair q, r: the proof above always produced a positive remainder r. If for example $b > 0$ and q, r are as above with $r > 0$, then $a = q'b + r'$ with $q' = q + 1$ and $r' = r - b$ also satisfy the conditions of the Division Algorithm applied to a, b. Thus $5 = 2 \cdot 2 + 1 = 3 \cdot 2 - 1$ are the two ways of applying the Division Algorithm in $\mathbb{Z}$ to $a = 5$ and $b = 2$. The quotient and remainder are unique if we require the remainder to be nonnegative.

(2) If F is a field, then the polynomial ring $F[x]$ is a Euclidean Domain with norm given by $N(p(x))$ = the degree of $p(x)$. The Division Algorithm for polynomials is simply "long division" of polynomials which may be familiar for polynomials with real coefficients. The proof is very similar to that for $\mathbb{Z}$ and is given in the next chapter (although for polynomials the quotient and remainder are shown to be unique). In order for a polynomial ring to be a Euclidean Domain the coefficients must come from a field since the division algorithm ultimately rests on being able to divide arbitrary nonzero coefficients. We shall prove in Section 2 that $R[x]$ is not a Euclidean Domain if R is not a field.

(3) The subring of the complex numbers $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ where $i^2 = -1$ is a Euclidean Domain with norm given by $N(a + bi) = a^2 + b^2$. This ring is called the *Gaussian integers* and its properties (including a proof of the Division Algorithm) are explored in the exercises.

(4) Recall (cf. Exercise 26 in Section 7.1) that a *discrete valuation ring* is obtained as follows. Let K be a field. A *discrete valuation* on K is a function $\nu : K^\times \to \mathbb{Z}$ satisfying

(i) $\nu(ab) = \nu(a) + \nu(b)$ (i.e. ν is a homomorphism from the multiplicative group of nonzero elements of K to $\mathbb{Z}$),

(ii) ν is surjective, and

(iii) $\nu(x+y) \geq \min\{\nu(x), \nu(y)\}$ for all $x, y \in K^\times$ with $x + y \neq 0$.

The set $\{x \in K^\times \mid \nu(x) \geq 0\} \cup \{0\}$ is a subring of K called the valuation ring of ν. An integral domain R is called a discrete valuation ring if there is a valuation ν on its field of fractions such that R is the valuation ring of ν.

For example the ring R of all rational numbers whose denominators are relatively prime to the fixed prime $p \in \mathbb{Z}$ is a discrete valuation ring contained in $\mathbb{Q}$.

A discrete valuation ring is easily seen to be a Euclidean Domain with respect to the norm defined by $N(0) = 0$ and $N = \nu$ on the nonzero elements of R. This is because for $a, b \in R$ with $b \neq 0$

(a) if $N(a) < N(b)$ then $a = 0 \cdot b + a$, and

(b) if $N(a) \geq N(b)$ then it follows from property (i) of a discrete valuation that $q = ab^{-1} \in R$, so $a = qb + 0$.

The first implication of a Division Algorithm for the integral domain R is that it forces every ideal of R to be *principal*.

Proposition 1. Every ideal in a Euclidean Domain is principal. More precisely, if I is any nonzero ideal in the Euclidean Domain R then $I = (d)$, where d is any nonzero element of I of minimum norm.

Proof: If I is the zero ideal, there is nothing to prove. Otherwise let d be any nonzero element of I of minimum norm (such a d exists since the set $\{N(a) \mid a \in I\}$ has a minimum element by the Well Ordering of $\mathbb{Z}$). Clearly $(d) \subseteq I$ since d is an element of I. To show the reverse inclusion let a be any element of I and use the Division Algorithm to write $a = qd + r$ with $r = 0$ or $N(r) < N(d)$. Then $r = a - qd$ and both a and qd are in I, so r is also an element of I. By the minimality of the norm of d, we see that r must be 0. Thus $a = qd \in (d)$ showing $I = (d)$.

Note that this now *proves* that every ideal of $\mathbb{Z}$ is principal. It also shows that $\mathbb{Z}[x]$ is *not* a Euclidean Domain (for any choice of norm) since the ideal $(2, x)$ is not principal in this ring (cf. Example 3 at the beginning of Section 7.4).

One of the fundamental consequences of the Euclidean Algorithm in $\mathbb{Z}$ is that it produces a greatest common divisor of two nonzero elements. This is true in any Euclidean Domain. The notion of a greatest common divisor of two elements (if it exists) can be made precise in general rings.

Definition. Let R be a commutative ring and let $a, b \in R$ with $b \neq 0$.

(1) a is said to be a *multiple* of b if there exists an element $x \in R$ with $a = bx$. In this case b is said to *divide* a or be a *divisor* of a, written $b \mid a$.

(2) A *greatest common divisor* of a and b is a nonzero element d such that

(i) $d \mid a$ and $d \mid b$, and

(ii) if $d' \mid a$ and $d' \mid b$ then $d' \mid d$.

A greatest common divisor of a and b will be denoted by g.c.d. (a, b), or (abusing the notation) simply (a, b).

Note that $b \mid a$ in a ring R if and only if $a \in (b)$ if and only if $(a) \subseteq (b)$. In particular, if d is any divisor of both a and b then (d) must contain both a and b and

hence must contain the ideal generated by a and b. The defining properties (i) and (ii) of a greatest common divisor of a and b translated into the language of ideals therefore become (respectively):

If I is the ideal of R generated by a and b, then d is a greatest common divisor of a and b if

(i) I is contained in the principal ideal (d), and

(ii) if (d') is any principal ideal containing I then $(d) \subseteq (d')$.

Thus a greatest common divisor of a and b (if such exists) is a generator for the unique smallest principal ideal containing a and b. There are rings in which greatest common divisors do not exist.

This discussion immediately gives the following *sufficient* condition for the existence of a greatest common divisor.

Proposition 2. If a and b are nonzero elements in the commutative ring R such that the ideal generated by a and b is a principal ideal (d), then d is a greatest common divisor of a and b.

This explains why the symbol (a, b) is often used to denote both the ideal generated by a and b and a greatest common divisor of a and b.

Note that the condition in Proposition 2 is *not* a *necessary* condition. For example, in the ring $R = \mathbb{Z}[x]$ the elements 2 and x generate a maximal, nonprincipal ideal (cf. the examples in Section 7.4). Thus $R = (1)$ is the unique principal ideal containing both 2 and x, so 1 is a greatest common divisor of 2 and x. We shall see other examples along these lines in Section 3.

Before returning to Euclidean Domains we examine the uniqueness of greatest common divisors.

Proposition 3. Let R be an integral domain. If two elements d and d' of R generate the same principal ideal, i.e. $(d) = (d')$, then $d' = ud$ for some unit u in R. In particular, if d and d' are both greatest common divisors of a and b, then $d' = ud$ for some unit u.

Proof: This is clear if either d or d' is zero so we may assume d and d' are nonzero. Since $d \in (d')$ there is some $x \in R$ such that $d = xd'$. Since $d' \in (d)$ there is some $y \in R$ such that $d' = yd$. Thus $d = xyd$ and so $d(1 - xy) = 0$. Since $d \neq 0$, $xy = 1$, that is, both x and y are units. This proves the first assertion. The second assertion follows from the first since any two greatest common divisors of a and b generate the same principal ideal (they divide each other).

One of the most important properties of Euclidean Domains is that greatest common divisors always exist and *can be computed algorithmically*.

Theorem 4. Let R be a Euclidean Domain and let a and b be nonzero elements of R. Let $d = r_n$ be the last nonzero remainder in the Euclidean Algorithm for a and b described at the beginning of this chapter. Then

(1) d is a greatest common divisor of a and b, and

(2) the principal ideal (d) is the ideal generated by a and b. In particular, d can be written as an *R-linear combination* of a and b, i.e. there are elements x and y in R such that

$$d = ax + by.$$

Proof: By Proposition 1, the ideal generated by a and b is principal so a, b do have a greatest common divisor, namely any element which generates the (principal) ideal (a, b). Both parts of the theorem will follow therefore once we show $d = r_n$ generates this ideal, i.e. once we show that

(i) $d \mid a$ and $d \mid b$ (so $(a, b) \subseteq (d)$)

(ii) d is an R-linear combination of a and b (so $(d) \subseteq (a, b)$).

To prove that d divides both a and b simply keep track of the divisibilities in the Euclidean Algorithm. Starting from the $(n+1)^{\text{st}}$ equation: $r_{n-1} = q_{n+1}r_n$ we see that $r_n \mid r_{n-1}$. Clearly $r_n \mid r_n$. By induction (proceeding from index n downwards to index 0) assume r_n divides r_{k+1} and r_k. By the $(k+1)^{\text{st}}$ equation: $r_{k-1} = q_{k+1}r_k + r_{k+1}$ and since r_n divides both terms on the right hand side we see that r_n also divides r_{k-1}. From the 1st equation in the Euclidean Algorithm we obtain that r_n divides b and then from the 0th equation we get that r_n divides a. Thus (i) holds.

To prove that r_n is in the ideal (a, b) generated by a and b proceed similarly by induction proceeding from equation (0) to equation (n). It follows from equation (0) that $r_0 \in (a, b)$ and by equation (1) that $r_1 = b - q_1 r_0 \in (b, r_0) \subseteq (a, b)$. By induction assume $r_{k-1}, r_k \in (a, b)$. Then by the $(k+1)^{\text{st}}$ equation:

$$r_{k+1} = r_{k-1} - q_{k+1}r_k \in (r_{k-1}, r_k) \subseteq (a, b).$$

This induction shows $r_n \in (a, b)$, which completes the proof.

Much of the material above may be familiar from elementary arithmetic in the case of the integers $\mathbb{Z}$, except possibly for the translation into the language of ideals. For example, if $a = 2210$ and $b = 1131$ then the smallest ideal of $\mathbb{Z}$ containing both a and b (the ideal generated by a and b) is $13\mathbb{Z}$, since 13 is the greatest common divisor of 2210 and 1131. This fact follows quickly from the Euclidean Algorithm:

$$\begin{aligned}
2210 &= 1 \cdot 1131 + 1079 \\
1131 &= 1 \cdot 1079 + 52 \\
1079 &= 20 \cdot 52 + 39 \\
52 &= 1 \cdot 39 + 13 \\
39 &= 3 \cdot 13
\end{aligned}$$

so that $13 = (2210, 1131)$ is the last nonzero remainder. Using the procedure of Theorem 4 we can also write 13 as a linear combination of 2210 and 1131 by first solving the next to last equation above for $13 = 52 - 1 \cdot 39$, then using previous equations to solve for 39 and 52, etc., finally writing 13 entirely in terms of 2210 and 1131. The answer in this case is:

$$13 = (-22) \cdot 2210 + 43 \cdot 1131.$$

The Euclidean Algorithm in the integers $\mathbb{Z}$ is extremely fast. It is a theorem that the number of steps required to determine the greatest common divisor of two integers a and b is at worst 5 times the number of digits of the smaller of the two numbers. Put another way, this algorithm is *logarithmic* in the size of the integers. To obtain an appreciation of the speed implied here, notice that for the example above we would have expected at worst $5 \cdot 4 = 20$ divisions (the example required far fewer). If we had started with integers on the order of 10^{100} (large numbers by physical standards), we would only have expected at worst 500 divisions.

There is no uniqueness statement for the integers x and y in $(a, b) = ax + by$. Indeed, $x' = x + b$ and $y' = y - a$ satisfy $(a, b) = ax' + by'$. This is essentially the only possibility — one can prove that if x_0 and y_0 are solutions to the equation $ax + by = N$, then any other solutions x and y to this equation are of the form

$$x = x_0 + m\frac{b}{(a, b)}$$
$$y = y_0 - m\frac{a}{(a, b)}$$

for some (positive or negative) integer m.

This latter theorem (a proof of which is outlined in the exercises) provides a complete solution of the *First Order Diophantine Equation* $ax + by = N$ provided we know there is *at least one* solution to this equation. But the equation $ax + by = N$ is simply another way of stating that N is an element of the ideal generated by a and b. Since we know this ideal is just (d), the principal ideal generated by the greatest common divisor d of a and b, this is the same as saying $N \in (d)$, i.e. N is divisible by d. Hence, *the equation $ax + by = N$ is solvable in integers x and y if and only if N is divisible by the g.c.d. of a and b* (and then the result quoted above gives a full set of solutions of this equation).

EXERCISES

1. For each of the following pairs of integers a and b, determine their greatest common divisor d and write d as a linear combination $ax + by$ of a and b.
   (a) $a = 20, b = 13$.
   (b) $a = 69, b = 372$.
   (c) $a = 11391, b = 5673$.
   (d) $a = 507885, b = 60808$.
   (e) $a = 91442056588823, b = 779086434385541$ (the Euclidean Algorithm requires only 7 steps for these integers).
2. For each of the following pairs of integers a and n, show that a is relatively prime to n and determine the inverse of a mod n (cf. Section 0.3).
   (a) $a = 13, n = 20$.
   (b) $a = 69, n = 89$.
   (c) $a = 1891, n = 3797$.
   (d) $a = 6003722857, n = 77695236973$ (the Euclidean Algorithm requires only 3 steps for these integers).
3. Let R be a Euclidean Domain. Let m be the minimum integer in the set of norms of nonzero elements of R. Prove that every nonzero element of R of norm m is a unit. Deduce that a nonzero element of norm zero (if such an element exists) is a unit.

4. Let R be a Euclidean Domain.
   (a) Prove that if $(a, b) = 1$ and a divides bc, then a divides c. More generally, show that if a divides bc with nonzero a, b then $\dfrac{a}{(a,b)}$ divides c.
   (b) Consider the Diophantine equation $ax + by = N$ where a, b and N are integers and a, b are nonzero. Suppose x_0, y_0 is a solution: $ax_0 + by_0 = N$. Prove that the full set of solutions to this equation is given by

$$x = x_0 + m\frac{b}{(a,b)}$$

$$y = y_0 - m\frac{a}{(a,b)}$$

   as m ranges over the integers. [If x, y is a solution to $ax+by = N$, show that $a(x-x_0) = b(y_0 - y)$ and use (a).]
5. Determine all integer solutions of the following equations:
   (a) $2x + 4y = 5$
   (b) $17x + 29y = 31$
   (c) $85x + 145y = 505$.
6. (*The Postage Stamp Problem*) Let a and b be two relatively prime positive integers. Prove that every sufficiently large positive integer N can be written as a linear combination $ax+by$ of a and b where x and y are both *positive*, i.e. there is an integer N_0 such that for all $N \geq N_0$ the equation $ax + by = N$ can be solved with both x and y positive integers. Prove in fact that the integer $ab - a - b$ cannot be written as a positive linear combination of a and b but that every integer greater than $ab - a - b$ is a positive linear combination of a and b (so every "postage" greater than $ab - a - b$ can be obtained using only stamps in denominations a and b).

For the next few exercises recall from Exercise 23 in Section 7.1 that for D an integer, not a perfect square

$$\mathbb{Z}[\sqrt{D}] = \{a + b\sqrt{D} \mid a, b \in \mathbb{Z}\}$$

is a subring of the complex numbers (and of the real numbers if $D > 0$). Recall also the map

$$N : \mathbb{Z}[\sqrt{D}] \to \mathbb{Z} \qquad \text{defined by} \qquad N(a + b\sqrt{D}) = a^2 - Db^2$$

which satisfies $N(\alpha\beta) = N(\alpha)N(\beta)$. Note that if $D > 0$ then N is not a norm (it is not nonnegative valued), although taking the absolute value of $N(\alpha)$ does give a norm. If $D < 0$ then N is a norm.

7. Take $D = -1$ and let $i = \sqrt{-1}$. Then $\mathbb{Z}[i]$ is called the ring of *Gaussian integers*. Prove that $\mathbb{Z}[i]$ is a Euclidean Domain with respect to N, hence is a Principal Ideal Domain, as follows.
   (a) Let α, β be two elements of $\mathbb{Z}[i]$ with $\beta \neq 0$. Show that $\dfrac{\alpha}{\beta} = r + si$ with *rational* numbers r, s. [Multiply numerator and denominator by $\overline{\beta}$, where bar denotes complex conjugation.]
   (b) Let p be an integer closest to the rational number r and let q be an integer closest to the rational number s. Let $\theta = (r - p) + (s - q)i$ and set $\gamma = \beta\theta$. Show that θ has norm at most $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and that γ is an element of $\mathbb{Z}[i]$ whose norm is at most $\frac{1}{2}N(\beta)$ [to show $\gamma \in \mathbb{Z}[i]$ observe that $\gamma = \alpha - (p + qi)\beta$]. Conclude that $\alpha = (p + qi)\beta + \gamma$ where $N(\gamma)$ is at most $\frac{1}{2}N(\beta)$. This gives an explicit Division Algorithm for $\mathbb{Z}[i]$.

(c) Use the procedure in (b) to find a generator for the ideal $(85, 1+13i)$, i.e. a greatest common divisor for 85 and $1 + 13i$, by the Euclidean Algorithm.

8. Prove that $\mathbb{Z}[\sqrt{-2}]$ is a Euclidean Domain with respect to N, hence is a Principal Ideal Domain, by the following steps that parallel the preceding exercise.

(a) Let α, β be two elements of $\mathbb{Z}[\sqrt{-2}]$, with $\beta \neq 0$. Show that $\dfrac{\alpha}{\beta} = r + s\sqrt{-2}$ with *rational* numbers r, s. [Multiply numerator and denominator by $\overline{\beta}$, where bar denotes complex conjugation.]

(b) Let p be an integer closest to the rational number r and let q be an integer closest to the rational number s. Let $\theta = (r - p) + (s - q)\sqrt{-2}$ and set $\gamma = \beta\theta$. Show that θ has norm at most $\dfrac{1}{4} + 2\dfrac{1}{4} = \dfrac{3}{4}$ and that γ is an element of $\mathbb{Z}[\sqrt{-2}]$ whose norm is at most $\dfrac{3}{4}N(\beta)$ [to show $\gamma \in \mathbb{Z}[\sqrt{-2}]$ observe that $\gamma = \alpha - (p + q\sqrt{-2})\beta$]. Conclude that $\alpha = (p + q\sqrt{-2})\beta + \gamma$ where $N(\gamma)$ is at most $\dfrac{3}{4}N(\beta)$. This gives an explicit Division Algorithm for $\mathbb{Z}[\sqrt{-2}]$.

(c) Use the procedure in (b) to find a generator for the ideal $(85, -11+4\sqrt{-2})$, i.e. a greatest common divisor for 85 and $-11 + 4\sqrt{-2}$, by the Euclidean Algorithm.

9. Prove that the quotient ring $\mathbb{Z}[i]/I$ is finite for any nonzero ideal I of $\mathbb{Z}[i]$. [Use the fact that $I = (\alpha)$ for some nonzero α and then use the Division Algorithm in this Euclidean Domain to see that every coset of I is represented by an element of norm less than $N(\alpha)$.]

10. Let R be a commutative ring with 1 and let a and b be nonzero elements of R. A *least common multiple* of a and b is an element e of R such that

(i) $a \mid e$ and $b \mid e$, and

(ii) if $a \mid e'$ and $b \mid e'$ then $e \mid e'$.

(a) Prove that a least common multiple of a and b (if such exists) is a generator for the unique largest principal ideal contained in $(a) \cap (b)$.

(b) Deduce that any two nonzero elements in a Euclidean Domain have a least common multiple which is unique up to multiplication by a unit.

(c) Prove that in a Euclidean Domain the least common multiple of a and b is $\dfrac{ab}{(a, b)}$, where (a, b) is the greatest common divisor of a and b.

11. (A Public Key Code) Let N be a positive integer. Let M be an integer relatively prime to N and let d be an integer relatively prime to $\varphi(N)$, where φ denotes Euler's φ-function. Prove that if $M' \equiv M^d \pmod{N}$ then $M \equiv M'^{d'} \pmod{N}$ where d' is the inverse of d mod $\varphi(N)$: $dd' \equiv 1 \pmod{\varphi(N)}$.

Remark: This result is the basis for a standard *Public Key Code*. Suppose $N = pq$ is the product of two distinct large primes (each on the order of 100 digits, for example). If M is a message, then $M' \equiv M^d \pmod{N}$ is a scrambled (*encoded*) version of M, which can be unscrambled (*decoded*) by computing $M'^{d'} \pmod{N}$ (these powers can be computed quite easily even for large values of M and N by successive squarings). The values of N and d (but not p and q) are made publicly known (hence the name) and then anyone with a message M can send their encoded message $M^d \pmod{N}$. To decode the message it seems necessary to determine d', which requires the determination of the value $\varphi(N) = \varphi(pq) = (p-1)(q-1)$ (no one has as yet *proved* that there is no other decoding scheme, however). The success of this method as a code rests on the necessity of determining the *factorization* of N into primes, for which no sufficiently efficient algorithm exists (for example, the most

naive method of checking all factors up to $\sqrt{N}$ would here require on the order of 10^{100} computations, or approximately 300 years even at 10 billion computations per second, and of course one can always increase the size of p and q).

8.2 PRINCIPAL IDEAL DOMAINS (P.I.D.s)

Definition. A *Principal Ideal Domain* (P.I.D.) is an integral domain in which every ideal is principal.

Proposition 1 proved that

$$\textit{every Euclidean Domain is a Principal Ideal Domain}$$

so that every result about Principal Ideal Domains automatically holds for Euclidean Domains. It is *not* true that every Principal Ideal Domain is a Euclidean Domain. Let θ be the complex number $\dfrac{1+\sqrt{-19}}{2}$; then the ring $\{a + b\theta \mid a, b \in \mathbb{Z}\}$ (which is a subring of $\mathbb{C}$) is a Principal Ideal Domain but not a Euclidean Domain (a proof is outlined in Exercises 5 to 8).

From an ideal-theoretic point of view Principal Ideal Domains are a natural class of rings to study beyond rings which are fields (where the ideals are just the trivial ones: (0) and (1)). Many of the properties enjoyed by Euclidean Domains are also satisfied by Principal Ideal Domains. A significant advantage of Euclidean Domains over Principal Ideal Domains, however, is that although greatest common divisors exist in both settings, in Euclidean Domains one has an *algorithm* for computing them. Thus (as we shall see in Chapter 12 in particular) results which depend on the existence of greatest common divisors may often be proved in the larger class of Principal Ideal Domains although computation of examples (i.e. concrete applications of these results) are more effectively carried out using a Euclidean Algorithm (if one is available).

We collect some facts about greatest common divisors proved in the preceding section.

Proposition 5. Let R be a Principal Ideal Domain and let a and b be nonzero elements of R. Let d be a generator for the principal ideal generated by a and b. Then:

(1) d is a greatest common divisor of a and b

(2) d can be written as an *R-linear combination* of a and b, i.e. there are elements x and y in R with

$$d = ax + by$$

(3) d is unique up to multiplication by a unit of R.

Proof: This is just Propositions 2 and 3.

Recall that maximal ideals are always prime ideals but the converse is not true in general. We observed in Section 7.4, however, that every nonzero prime ideal of $\mathbb{Z}$ is a maximal ideal. This useful fact is true in an arbitrary Principal Ideal Domain.

Proposition 6. Every nonzero prime ideal in a Principal Ideal Domain is a maximal ideal.

Proof: Let (p) be a nonzero prime ideal in the Principal Ideal Domain R. By Proposition 11 in Section 7.4 there is a maximal ideal M containing (p). Since R is a Principal Ideal Domain, $M = (m)$ for some $m \in R$. Now $p \in (m)$ so $p = rm$ for some $r \in R$. Since (p) is a prime ideal and $rm \in (p)$, either r or m must lie in (p). If $m \in (p)$ then $(p) = (m)$ is maximal and the proof is complete. If, on the other hand, $r \in (p)$ write $r = ps$. In this case $p = rm = psm$, so $sm = 1$ (recall that R is an integral domain) and m is a unit. This is a contradiction since (m) is, by choice, a proper ideal.

As we have already mentioned, if F is a field, then the polynomial ring $F[x]$ is a Euclidean Domain, hence also a Principal Ideal Domain (this will be proved in the next chapter). The converse to this is also true. Intuitively, if I is an ideal in R (such as the ideal (2) in $\mathbb{Z}$) then the ideal (I, x) in $R[x]$ (such as the ideal $(2, x)$ in $\mathbb{Z}[x]$) requires one more generator than does I, hence in general is not principal.

Corollary 7. If R is any commutative ring such that the polynomial ring $R[x]$ is a Principal Ideal Domain (or a Euclidean Domain), then R is necessarily a field.

Proof: Assume $R[x]$ is a Principal Ideal Domain. Since R is a subring of $R[x]$ then R must be an integral domain (recall that $R[x]$ has an identity if and only if R does). The ideal (x) is a nonzero prime ideal in $R[x]$ because $R[x]/(x)$ is isomorphic to the integral domain R. By Proposition 6, (x) is a maximal ideal, hence the quotient R is a field by Proposition 12 in Section 7.4.

EXERCISES

1. Prove that in a Principal Ideal Domain two ideals (a) and (b) are comaximal (cf. Section 7.6) if and only if a greatest common divisor of a and b is 1.
2. Prove that any two nonzero elements of a P.I.D. have a least common multiple (cf. Exercise 10, Section 1).
3. Let R be an integral domain. Prove that if the following two conditions hold then R is a Principal Ideal Domain:
   (i) any two nonzero elements a and b in R have a greatest common divisor which can be written in the form $ra + sb$ for some $r, s \in R$, and
   (ii) if $a_1, a_2, a_3, \ldots$ are nonzero elements of R such that $a_{i+1} \mid a_i$ for all i, then there is a positive integer N such that a_n is a unit times a_N for all $n \geq N$.
4. Prove that if R is a Principal Ideal Domain and D is a multiplicatively closed subset of R, then $D^{-1}R$ is also a P.I.D. (cf. Section 7.5).

The following exercises outline a proof that the set of elements of the form $a + b\left(1 + \sqrt{-19}/2\right)$ with $a, b \in \mathbb{Z}$ is a Principal Ideal Domain that is not a Euclidean Domain. These exercises follow the exposition by J.C. Wilson[1] of ideas of Th. Motzkin, and use a simplification by Kenneth S. Williams[2].

[1] In *A principal ideal ring that is not a Euclidean ring*, Math. Mag., 46(1973), pp. 34-38

[2] In *Note on non-Euclidean Principal Ideal Domains*, Math. Mag., 48(1975), pp. 176-177.

5. Let

$$F = \{A + B\sqrt{-19} \mid A, B \in \mathbb{Q}\} \subset \mathbb{C}$$

and let

$$R = \{a + b\left(\frac{1+\sqrt{-19}}{2}\right) \mid a, b \in \mathbb{Z}\} \subset F .$$

(a) Prove that F is a field and that R is a ring, hence is an integral domain. Prove that F is the field of fractions of R.

(b) If $\alpha = A + B\sqrt{-19} \in F$, let $\bar{\alpha} = A - B\sqrt{-19} \in F$ and set $N(\alpha) = \alpha\bar{\alpha} = A^2 + 19B^2$. Prove that $N(\alpha) > 0$ for $\alpha \neq 0$ and that N is multiplicative: $N(\alpha\beta) = N(\alpha)N(\beta)$ for all $\alpha, \beta \in F$. Finally, prove that $N(\alpha)$ is a positive integer for all nonzero $\alpha \in R$.

(c) Prove that ± 1 are the only units in R. Prove that the only factors of 2 in R are $\{\pm 1, \pm 2\}$ and the only factors of 3 in R are $\{\pm 1, \pm 3\}$.

6. (*Criterion of Dedekind and Hasse*) Let S be any integral domain and let N denote any function from S to $\mathbb{Z}$ which satisfies $N(\alpha) > 0$ for $\alpha \neq 0$. Suppose that for every $\alpha, \beta \in S$ with $N(\alpha) > N(\beta)$ either β divides α in S or there exist $s, t \in S$ with $0 < N(s\alpha - t\beta) < N(\beta)$. Prove that S is a Principal Ideal Domain. (Note that S is Euclidean with respect to N if it is always possible to satisfy this condition with $s = 1$, so this is a weakening of the Euclidean condition.) [Let I be any nonzero ideal in S and let β be a nonzero element of I with $N(\beta)$ minimal. If $\alpha \in I$ then $s\alpha - t\beta$ is also in I for all $s, t \in S$. Use minimality to conclude that $\alpha \in (\beta)$, so $I = (\beta)$ is principal.]

7. This exercise shows that the function N defined on the ring R in Exercise 5 satisfies the properties in the previous exercise, proving that R is a Principal Ideal Domain. Suppose α, β are nonzero elements of R and $\alpha/\beta \notin R$. Note that since N is multiplicative the condition $0 < N(s\alpha - t\beta) < N(\beta)$ is equivalent to

$$0 < N(\frac{\alpha}{\beta}s - t) < 1 . \qquad (*)$$

Write $\dfrac{\alpha}{\beta} = \dfrac{a + b\sqrt{-19}}{c} \in F$ with integers a, b, c having no common divisor and with $c > 1$ (since β is assumed not to divide α).

(a) Show that since a, b, c have no common divisor that there are integers x, y, z with $ax + by + cz = 1$. Write $ay - 19bx = cq + r$ for some quotient q and remainder r with $|r| \leq c/2$. Show that $s = y + x\sqrt{-19}$ and $t = q - z\sqrt{-19}$ satisfy $(*)$ if $c \geq 5$.

(b) Suppose $c = 2$. Show that one of a, b is even and the other is odd and that $s = 1$, $t = \dfrac{(a-1) + b\sqrt{-19}}{2}$ are elements of R that satisfy $(*)$.

(c) Suppose $c = 3$. Show that $a^2 + 19b^2$ is not divisible by 3. Write $a^2 + 19b^2 = 3q + r$ with $r = 1$ or 2 and show that $s = a - b\sqrt{-19}$, $t = q$ satisfy $(*)$.

(d) Suppose $c = 4$, so a and b are not both even. If one of a, b is even, the other odd, show that $a^2 + 19b^2 = 4q + r$ for some $q, r \in \mathbb{Z}$ and $0 < r < 4$ and that $s = a - b\sqrt{-19}$, $t = q$ satisfy $(*)$. Finally, if a and b are both odd, show that $a^2 + 19b^2 = 8q + 4$ for some $q \in \mathbb{Z}$ and that $s = \dfrac{a - b\sqrt{-19}}{2}$, $t = q$ are elements of R that satisfy $(*)$.

8. This exercise shows that the Principal Ideal Domain R above is not a Euclidean Domain. Let D be any integral domain and let $\widetilde{D}$ denote the collection of units of D together with 0. An element $u \in D - \widetilde{D}$ is called a *universal side divisor* if for every $x \in D$ there is some $z \in \widetilde{D}$ such that u divides $x - z$ in D.

(a) Prove that if D is not a field and D has no universal side divisors then D is not a Euclidean Domain. (So the existence of universal side divisors is also a weakening of the

Euclidean condition.) [Suppose D is Euclidean with respect to some norm and let u be an element of $D - \widetilde{D}$ of minimal norm. For every $x \in D$, write $x = qu + r$ and use minimality to show that $r = 0$ or $r \in \widetilde{D}$, showing that u is a universal side divisor.]

(b) Prove that the ring R above has no universal side divisors, hence is not a Euclidean Domain. [Suppose u is a universal side divisor. Show $\widetilde{R} = \{0, \pm 1\}$. Let $x = 2$ and conclude that u is a nonunit divisor of 2 or 3, hence is ± 2 or ± 3 by (c) in Exercise 5. Let $x = \dfrac{1 + \sqrt{-19}}{2}$ to show that none of these is a universal side divisor.]

8.3 UNIQUE FACTORIZATION DOMAINS (U.F.D.s)

In the case of the integers $\mathbb{Z}$, there is another method for determining the greatest common divisor of two elements a and b familiar from elementary arithmetic, namely the notion of "factorization into primes" for a and b, from which the greatest common divisor can easily be determined. This can also be extended to a larger class of rings called Unique Factorization Domains (U.F.D.s) — these will be defined shortly. We shall then prove that

every Principal Ideal Domain is a Unique Factorization Domain

so that every result about Unique Factorization Domains will automatically hold for both Euclidean Domains and Principal Ideal Domains.

We first introduce some terminology.

Definition. Let R be an integral domain.

(1) Suppose $r \in R$ is nonzero and is not a unit. Then r is called *irreducible* in R if whenever $r = ab$ with $a, b \in R$, at least one of a or b must be a unit in R. Otherwise r is said to be *reducible*.

(2) The nonzero element $p \in R$ is called *prime* in R if the ideal (p) generated by p is a prime ideal. In other words, a nonzero element p is a prime if it is not a unit and whenever $p \mid ab$ for any $a, b \in R$, then either $p \mid a$ or $p \mid b$.

(3) Two elements a and b of R differing by a unit are said to be *associate* in R (i.e. $a = ub$ for some unit u in R).

Proposition 8. In an integral domain a prime element is always irreducible.

Proof: Suppose (p) is a nonzero prime ideal and $p = ab$. Then $ab = p \in (p)$, so by definition of prime ideal one of a or b, say a, is in (p). Thus $a = pr$ for some r. This implies $p = ab = prb$ so $rb = 1$ and b is a unit. This shows that p is irreducible.

It is not true in general that an irreducible element is necessarily prime. For example, consider the subring of the complex numbers $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$. The element $\alpha = 2 + \sqrt{-5}$ is irreducible (cf. the exercises) but is not a prime since $3^2 = (2 + \sqrt{-5})(2 - \sqrt{-5})$ so $3^2 \in (\alpha)$ but $3 \notin (\alpha)$.

If R is a Principal Ideal Domain however, the notions of prime and irreducible elements are the same. In particular these notions coincide in $\mathbb{Z}$ and in $F[x]$ (where F is a field).

Proposition 9. In a Principal Ideal Domain a nonzero element is a prime if and only if it is irreducible.

Proof: We have shown above that prime implies irreducible. We must show conversely that if p is irreducible, then p is a prime, i.e. the ideal (p) is a prime ideal. If M is any ideal containing (p) then by hypothesis $M = (m)$ is a principal ideal. Since $p \in (m)$, $p = rm$ for some r. But p is irreducible so by definition either r or m is a unit. This means either $(p) = (m)$ or $(m) = (1)$, respectively. Thus the only ideals containing (p) are (p) or (1), i.e. (p) is a maximal ideal. Since maximal ideals are prime ideals, the proof is complete.

The irreducible elements in the integers $\mathbb{Z}$ are the prime numbers (and their negatives) familiar from elementary arithmetic, and two integers a and b are associates of one another if and only if $a = \pm b$.

In the integers $\mathbb{Z}$ any integer n can be written as a product of primes (not necessarily distinct), as follows. If n is not itself a prime then by definition it is possible to write $n = n_1 n_2$ for two other integers n_1 and n_2 neither of which is a unit, i.e. neither of which is ± 1. Both n_1 and n_2 must be smaller in absolute value than n itself. If they are both primes, we have already written n as a product of primes. If one of n_1 or n_2 is not prime, then it in turn can be factored into two (smaller) integers. Since integers cannot decrease in absolute value indefinitely, we must at some point be left only with prime integer factors, and so we have written n as a product of primes.

For example, if $n = 2210$, the algorithm above proceeds as follows: n is not itself prime, since we can write $n = 2 \cdot 1105$. The integer 2 is a prime, but 1105 is not: $1105 = 5 \cdot 221$. The integer 5 is prime, but 221 is not: $221 = 13 \cdot 17$. Here the algorithm terminates, since both 13 and 17 are primes. This gives the *prime factorization* of 2210 as $2210 = 2 \cdot 5 \cdot 13 \cdot 17$. Similarly, we find $1131 = 3 \cdot 13 \cdot 29$. In these examples each prime occurs only to the first power, but of course this need not be the case generally.

In the ring $\mathbb{Z}$ not only is it true that every integer n can be written as a product of primes, but in fact this decomposition is *unique* in the sense that any two prime factorizations of the same positive integer n differ only in the order in which the positive prime factors are written. The restriction to positive integers is to avoid considering the factorizations $(3)(5)$ and $(-3)(-5)$ of 15 as essentially distinct. This *unique factorization* property of $\mathbb{Z}$ (which we shall prove very shortly) is extremely useful for the arithmetic of the integers. General rings with the analogous property are given a name.

Definition. A *Unique Factorization Domain* (U.F.D.) is an integral domain R in which every nonzero element $r \in R$ which is not a unit has the following two properties:

(i) r can be written as a finite product of irreducibles p_i of R (not necessarily distinct): $r = p_1 p_2 \ldots p_n$ and

(ii) the decomposition in (i) is *unique up to associates*: namely, if $r = q_1 q_2 \ldots q_m$ is another factorization of r into irreducibles, then $m = n$ and there is some renumbering of the factors so that p_i is associate to q_i for $i = 1, 2, \ldots, n$.

Examples

(1) A field F is trivially a Unique Factorization Domain since every nonzero element is a unit, so there are no elements for which properties (i) and (ii) must be verified.

(2) As indicated above, we shall prove shortly that every Principal Ideal Domain is a Unique Factorization Domain (so, in particular, $\mathbb{Z}$ and $F[x]$ where F is a field are both Unique Factorization Domains).

(3) We shall also prove in the next chapter that the ring $R[x]$ of polynomials is a Unique Factorization Domain whenever R itself is a Unique Factorization Domain (in contrast to the properties of being a Principal Ideal Domain or being a Euclidean Domain, which do not carry over from a ring R to the polynomial ring $R[x]$). This result together with the preceding example will show that $\mathbb{Z}[x]$ is a Unique Factorization Domain.

(4) The subring of the complex numbers: $R = \mathbb{Z}[2i] = \{a+2bi \mid a, b \in \mathbb{Z}\}$, where $i^2 = -1$, is an integral domain but not a Unique Factorization Domain. The elements 2 and $2i$ are irreducibles which are not associates in R (since $i \notin R$) and $4 = 2 \cdot 2 = (-2i)\cdot(2i)$ has two distinct factorizations in R. In the larger ring of Gaussian integers, $\mathbb{Z}[i]$, (which is a Unique Factorization Domain) 2 and $2i$ *are* associates since i is a unit in this larger ring (but i does not belong to its subring R). We shall give a slightly different proof that $\mathbb{Z}[2i]$ is not a Unique Factorization Domain at the end of Section 9.3 (one in which we do not have to check that 2 and $2i$ are irreducibles).

The first property of irreducible elements in a Unique Factorization Domain is that they are also primes. One might think that we could deduce Proposition 9 from this proposition together with the theorem that every Principal Ideal Domain is a Unique Factorization Domain, however Proposition 9 will be used in the proof of the latter theorem.

Proposition 10. In a Unique Factorization Domain a nonzero element is a prime if and only if it is irreducible.

Proof: Let R be a Unique Factorization Domain. Since by Proposition 8, primes of R are irreducible it remains to prove that each irreducible element is a prime. Let p be an irreducible in R and assume $p \mid ab$ for some $a, b \in R$; we must show that p divides either a or b. To say that p divides ab is to say $ab = pc$ for some c in R. Writing a and b as a product of irreducibles, we see from this last equation and from the *uniqueness* of the decomposition into irreducibles of ab that the irreducible element p must be *associate* to one of the irreducibles occurring either in the factorization of a or in the factorization of b. We may assume that p is associate to one of the irreducibles in the factorization of a, i.e. that a can be written as a product $a = (up)p_2 \dots p_n$ for u a unit and some (possibly empty set of) irreducibles $p_2, \dots, p_n$. But then p divides a, since $a = pd$ with $d = up_2 \dots p_n$, completing the proof.

Example

Proposition 10 shows that $\mathbb{Z}[\sqrt{-5}]$ is another example of an integral domain that is not a Unique Factorization Domain — recall that $2 + \sqrt{-5}$ is irreducible but not prime in this ring.

In a Unique Factorization Domain we shall now use the terms "prime" and "irreducible" interchangeably although we shall usually refer to the "primes" in $\mathbb{Z}$ and the "irreducibles" in $F[x]$.

We shall use the preceding proposition to show that in a Unique Factorization Domain any two nonzero elements a and b have a greatest common divisor:

Proposition 11. Let a and b be two nonzero elements of the Unique Factorization Domain R and suppose

$$a = up_1{}^{e_1}p_2{}^{e_2}\dots p_n{}^{e_n} \qquad \text{and} \qquad b = vp_1{}^{f_1}p_2{}^{f_2}\dots p_n{}^{f_n}$$

are prime factorizations for a and b, where u and v are units, the primes $p_1, p_2, \dots, p_n$ are *distinct* and the exponents e_i and f_i are ≥ 0. Then the element

$$d = p_1{}^{\min(e_1,f_1)}p_2{}^{\min(e_2,f_2)}\dots p_n{}^{\min(e_n,f_n)}$$

(where $d = 1$ if all the exponents are 0) is a greatest common divisor of a and b.

Proof: Since the exponents of each of the primes occurring in d are no larger than the exponents occurring in the factorizations of both a and b, d divides both a and b. To show that d is a greatest common divisor, let c be any common divisor of a and b and let $c = q_1{}^{g_1}q_2{}^{g_2}\dots q_m{}^{g_m}$ be the prime factorization of c. Since each q_i divides c, hence divides a and b, we see from the preceding proposition that q_i must divide one of the primes p_j. In particular, up to associates (so up to multiplication by a unit) the primes occurring in c must be a subset of the primes occurring in a and b : $\{q_1, q_2, \dots, q_m\} \subseteq \{p_1, p_2, \dots, p_n\}$. Similarly, the exponents for the primes occurring in c must be no larger than those occurring in d. This implies that c divides d, completing the proof.

Example

In the example above, where $a = 2210$ and $b = 1131$, we find immediately from their prime factorizations that $(a, b) = 13$. Note that if the prime factorizations for a and b are known, the proposition above gives their greatest common divisor instantly, but that finding these prime factorizations is extremely time-consuming computationally. The Euclidean Algorithm is the fastest method for determining the g.c.d. of two integers but unfortunately it gives almost no information on the prime factorizations of the integers.

We now come to one of the principal results relating some of the rings introduced in this chapter.

Theorem 12. Every Principal Ideal Domain is a Unique Factorization Domain. In particular, every Euclidean Domain is a Unique Factorization Domain.

Proof: Note that the second assertion follows from the first since Euclidean Domains are Principal Ideal Domains. To prove the first assertion let R be a Principal Ideal Domain and let r be a nonzero element of R which is not a unit. We must show first that r can be written as a finite product of irreducible elements of R and then we must verify that this decomposition is unique up to units.

The method of proof of the first part is precisely analogous to the determination of the prime factor decomposition of an integer. Assume r is nonzero and is not a unit. If r is itself irreducible, then we are done. If not, then by definition r can be written as a product $r = r_1 r_2$ where neither r_1 nor r_2 is a unit. If both these elements are irreducibles, then again we are done, having written r as a product of irreducible elements. Otherwise, at least one of the two elements, say r_1 is reducible, hence can be written as a product of two nonunit elements $r_1 = r_{11} r_{12}$, and so forth. What we must verify is that this process *terminates*, i.e. that we must necessarily reach a point where all of the elements obtained as factors of r are irreducible. Suppose this is not the case. From the factorization $r = r_1 r_2$ we obtain a *proper* inclusion of ideals: $(r) \subset (r_1) \subset R$. The first inclusion is proper since r_2 is not a unit, and the last inclusion is proper since r_1 is not a unit. From the factorization of r_1 we similarly obtain $(r) \subset (r_1) \subset (r_{11}) \subset R$. If this process of factorization did not terminate after a finite number of steps, then we would obtain an *infinite ascending chain* of ideals:

$$(r) \subset (r_1) \subset (r_{11}) \subset \cdots \subset R$$

(where all containments are proper).

We now show that any ascending chain $I_1 \subseteq I_2 \subseteq \cdots \subseteq R$ of ideals in a Principal Ideal Domain eventually becomes stationary, i.e. there is some positive integer n such that $I_k = I_n$ for all $k \geq n$.[3] In particular, it is not possible to have an infinite ascending chain of ideals where all containments are proper. Let $I = \cup_{i=1}^{\infty} I_i$. It follows easily (as in the proof of Proposition 11 in Section 7.4) that I is an ideal. Since R is a Principal Ideal Domain it is principally generated, say $I = (a)$. Since I is the union of the ideals above, a must be an element of one of the ideals in the chain, say $a \in I_n$. But then we have $I_n \subseteq I = (a) \subseteq I_n$ and so $I = I_n$ and the chain becomes stationary at I_n. This proves that every nonzero element of R which is not a unit has some factorization into irreducibles in R.

It remains to prove that the above decomposition is essentially unique. We proceed by induction on the number, n, of irreducible factors in some factorization of the element r. If $n = 0$, then r is a unit. If we had $r = qc$ (some other factorization) for some irreducible q, then q would divide a unit, hence would itself be a unit, a contradiction. Suppose now that n is at least 1 and that we have two products

$$r = p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m \qquad m \geq n$$

for r where the p_i and q_j are (not necessarily distinct) irreducibles. Since then p_1 divides the product on the right, we see by Proposition 10 that p_1 must divide one of the factors. Renumbering if necessary, we may assume p_1 divides q_1. But then $q_1 = p_1 u$ for some element u of R which must in fact be a unit since q_1 is irreducible. Thus p_1 and q_1 are associates. Cancelling p_1 (recall we are in an integral domain, so this is legitimate), we obtain the equation

$$p_2 \cdots p_n = u q_2 q_3 \cdots q_m = q_2' q_3 \cdots q_m \qquad m \geq n.$$

[3] This same argument can be used to prove the more general statement: an ascending chain of ideals becomes stationary in any commutative ring where all the ideals are *finitely generated*. This result will be needed in Chapter 12 where the details will be repeated.

where $q_2' = uq_2$ is again an irreducible (associate to q_2). By induction on n, we conclude that each of the factors on the left matches bijectively (up to associates) with the factors on the far right, hence with the factors in the middle (which are the same, up to associates). Since p_1 and q_1 (after the initial renumbering) have already been shown to be associate, this completes the induction step and the proof of the theorem.

Corollary 13. (The Fundamental Theorem of Arithmetic) The integers $\mathbb{Z}$ are a Unique Factorization Domain.

Proof: The integers $\mathbb{Z}$ are a Euclidean Domain, hence are a Unique Factorization Domain by the theorem.

Summary

We have the following inclusions among classes of commutative rings with identity:

$$fields \subset Euclidean\ Domains \subset P.I.D.s \subset U.F.D.s \subset integral\ domains$$

with all containments being proper. Recall that $\mathbb{Z}$ is a Euclidean Domain that is not a field, $\mathbb{Z}[\theta]$, where $\theta = \dfrac{1+\sqrt{-19}}{2}$ is a Principal Ideal Domain that is not a Euclidean Domain (a proof of this is given in the Exercises of Section 2), $\mathbb{Z}[x]$ is a Unique Factorization Domain that is not a Principal Ideal Domain (Theorem 7 in Chapter 9) and $\mathbb{Z}[2i]$ is an integral domain that is not a Unique Factorization Domain (proved at the end of Section 9.3).

EXERCISES

Let D be an integer, not a perfect square and recall the ring $\mathbb{Z}[\sqrt{D}] = \{a + b\sqrt{D} \mid a, b \in \mathbb{Z}\}$ introduced in Exercise 23 in Section 7.1 and explored further (for certain values of D) in the exercises in Section 1. Recall also the map $N(a + b\sqrt{D}) = a^2 - Db^2$ which satisfies $N(\alpha\beta) = N(\alpha)N(\beta)$.

1. Show that if $N(\alpha)$ is $\pm$ a prime (in $\mathbb{Z}$), then α is irreducible in $\mathbb{Z}[\sqrt{D}]$. [Recall that β is a unit in $\mathbb{Z}[\sqrt{D}]$ if and only if $N(\beta) = \pm 1$.]
2. Specialize to the case $D = -1$ above, so $\mathbb{Z}[\sqrt{-1}] = \mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ is the ring of *Gaussian integers* and $N(\alpha) = \alpha\overline{\alpha} = a^2 + b^2$ in this case.
   (a) Determine the units in $\mathbb{Z}[i]$.
   (b) Show that $2 = (1+i)(1-i)$ and $5 = (2+i)(2-i)$ in $\mathbb{Z}[i]$, where $1+i$, $1-i$, $2+i$, $2-i$ are irreducibles in $\mathbb{Z}[i]$. Show that 3 and 7 are irreducible in $\mathbb{Z}[i]$.
   (c) Determine which of the following are irreducible in $\mathbb{Z}[i]$: 11, 13, 17, 19, 23, 29, 31.
3. Specialize to the case $D = -2$ above, so $\mathbb{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} \mid a, b \in \mathbb{Z}\}$ and $N(\alpha) = \alpha\overline{\alpha} = a^2 + 2b^2$ in this case.
   (a) Show that the only units in $\mathbb{Z}[\sqrt{-2}]$ are ± 1.
   (b) Show that $2 = -(\sqrt{-2})^2$ and $3 = (1+\sqrt{-2})(1-\sqrt{-2})$ in $\mathbb{Z}[\sqrt{-2}]$, where $\sqrt{-2}$, $1+\sqrt{-2}$, $1-\sqrt{-2}$ are irreducibles in $\mathbb{Z}[\sqrt{-2}]$. Show that 5 and 7 are irreducible in $\mathbb{Z}[\sqrt{-2}]$.
   (c) Determine which of the following are irreducible in $\mathbb{Z}[\sqrt{-2}]$: 11, 13, 17, 19, 23, 29, 31.
4. Specialize to the case $D = -5$ above, so $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$ and $N(\alpha) = \alpha\overline{\alpha} = a^2 + 5b^2$ in this case.

(a) Prove that $\alpha = 2+\sqrt{-5}$ is irreducible in $\mathbb{Z}[\sqrt{-5}]$ but is not prime. [If $\alpha = \beta\gamma$, consider $N(\alpha) = N(\beta)N(\gamma)$. The fact that α is not prime was indicated in the text.]

(b) Conclude from (a) that R is *not* a P.I.D., so in particular is not Euclidean (with respect to *any* norm). Show in fact that the ideal $(3, 2+\sqrt{-5})$ is not a principal ideal.

5. Specialize to the case $D = 2$ above, so $\mathbb{Z}[\sqrt{2}] = \{a+b\sqrt{2} \mid a, b \in \mathbb{Z}\}$ and $N(\alpha) = a^2-2b^2$ in this case.

(a) Show that the element $\epsilon = 1 + \sqrt{2}$ is a unit in $\mathbb{Z}[\sqrt{2}]$. Show that the elements $\pm\epsilon^n$ are units for any integer n, so this ring contains infinitely many units (in fact it is true that these are the only units in $\mathbb{Z}[\sqrt{2}]$, but this is harder to prove). Prove that there are infinitely many solutions in integers x, y to the equations $x^2 - 2y^2 = 1$ and $x^2 - 2y^2 = -1$ (these are examples of *Pell's equation*: $x^2 - Dy^2 = \pm 1$ with $D = 2$).

(b) Prove that $\mathbb{Z}[\sqrt{2}]$ is a Euclidean Domain with respect to the norm $\alpha \mapsto |N(\alpha)|$ [Follow the steps of Exercise 8, Section 1.]

6. Let $G = \mathbb{Q}^\times$ be the multiplicative group of nonzero rational numbers. If $\alpha = \dfrac{p}{q} \in G$, where p and q are relatively prime integers, let $\varphi : G \to G$ be the map which interchanges the primes 2 and 3 in the prime power factorizations of p and q (so, for example, $\varphi(2^4 3^{11} 5^1 13^2) = 3^4 2^{11} 5^1 13^2$, $\varphi(\frac{3}{16}) = \varphi(\frac{3}{2^4}) = \frac{2}{3^4} = \frac{2}{81}$, and φ is the identity on all rational numbers whose numerators and denominators are relatively prime to 2 and 3).

(a) Prove that φ is a group isomorphism.

(b) Prove that there are infinitely many isomorphisms of the group G to itself.

(c) Prove that none of the isomorphisms above can be extended to an isomorphism of the *ring* $\mathbb{Q}$ to itself. In fact prove that the identity map is the only ring isomorphism of $\mathbb{Q}$.

7. Let a and b be nonzero elements of the Unique Factorization Domain R. Prove that a and b have a least common multiple (cf. Exercise 10 of Section 1) and describe it in terms of the prime factorizations of a and b in the same fashion that Proposition 11 describes their greatest common divisor.

The next few exercises explore a natural generalization of the class of rings in which every ideal is principal. Let R be a commutative ring with 1. Say R is a *Noetherian* ring[4] if every ideal of R is finitely generated, i.e. $I = (a_1, a_2, \ldots, a_n)$ for every ideal I of R (where n may depend on I). In particular, every Principal Ideal Domain is a Noetherian ring. Noetherian rings are by far the most important class of rings in commutative algebra. The first exercise provides some characterizations of Noetherian rings.

8. Prove that the following three conditions are equivalent:

(i) R is a Noetherian ring

(ii) if $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$ is any ascending chain of ideals of R, then there is a positive integer N such that $I_k = I_N$ for all $k \geq N$, i.e. the chain becomes stationary at I_N (if this holds, R is said to satisfy the *ascending chain condition* on ideals)

(iii) every nonempty collection of ideals of R contains a maximal element under inclusion.

[Show (i) $\Rightarrow$ (ii) by arguing as in the proof of Theorem 12. Show (ii) $\Rightarrow$ (iii) by assuming (iii) fails and constructing an infinite, nonstationary chain of ideals. Show (iii) $\Rightarrow$ (i) by letting I be any ideal and considering the collection of all finitely generated ideals that are contained in I.]

9. Prove that any quotient of a Noetherian ring is also a Noetherian ring.

10. Prove that if R is a Noetherian ring and D is a nonempty, multiplicatively closed subset of R then $D^{-1}R$ is a Noetherian ring (cf. Section 7.5).

[4] Named after Emmy Noether.

11. This exercise outlines a proof of *Hilbert's Basis Theorem*: if R is a Noetherian ring then so is the polynomial ring $R[x]$.

Let R be a Noetherian ring, let x be an indeterminate over R and let I be an ideal in the polynomial ring $R[x]$. The aim is to show that I is finitely generated. Since the zero ideal is finitely generated, assume $I \neq 0$.

(a) Recall that if $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ with $a_n \neq 0$, then a_n is called the *leading coefficient* of f (with the convention that the leading coefficient of the zero polynomial is 0). Let L be the subset of R consisting of all leading coefficients of all polynomials in I. Prove that L is an ideal of R.

(b) Since R is Noetherian, L is generated by a finite set. Pick polynomials $f_1, f_2, \ldots, f_m$ in I such that the leading coefficients of these polynomials generate L. Fix an integer N that is greater than the degree of f_i for all i. For each $j \in \{1, 2, \ldots, N\}$ let L_j be the subset of R consisting of all leading coefficients of all polynomials in I of degree $\leq j$. Prove that L_j is an ideal of R.

(c) Since R is Noetherian, L_j is generated by a finite set for each $j \in \{1, 2, \ldots, N\}$. Pick polynomials $f_{1j}, f_{2j}, \ldots, f_{kj}$ in I such that the leading coefficients of these polynomials generate L_j (where k depends on j). Let I' be the ideal of $R[x]$ generated by the finite set of polynomials: $f_1, \ldots, f_m$ and f_{ij} for all i, j. Note that $I' \subseteq I$. Complete the proof of Hilbert's Basis Theorem by showing $I' = I$ as follows. Let g be an element of I of smallest degree subject to $g \notin I'$ (if no such g exists, $I = I'$ is finitely generated). If $\deg g > N$, show that there are elements $p_1, \ldots, p_m \in R[x]$ such that $g - (p_1 f_1 + \cdots + p_m f_m)$ is an element of $I - I'$ of smaller degree than g (contrary to the minimality of g). If, on the other hand, $\deg g = j \leq N$, show that one may similarly subtract some $R[x]$-linear combination of the polynomials f_{ij} from g to obtain another element of $I - I'$ of smaller degree than g (giving the final contradiction).

12. Prove that if R is a Noetherian ring then so is the ring of formal power series $R[[x]]$ (cf. Exercise 3, Section 7.2). [Follow the same steps as in the preceding proof but use the coefficients of the terms of *lowest* degree of elements in an ideal I of $R[[x]]$.]

13. Let R be a Noetherian ring and let I be an ideal of R. Prove that I contains some power of $\mathcal{R}(I)$, where $\mathcal{R}(I)$ is the radical of I defined in Exercise 30 of Section 7.4. Deduce that some power of the nilradical is 0, i.e. the nilradical of a Noetherian ring is a nilpotent ideal (a power of an ideal is defined in Exercise 36 of Section 7.3 and nilpotent ideals are defined in Exercise 37 of the same section). [See Exercise 28, Section 7.4.]

14. (*k–stage Euclidean Domains*) Let R be an integral domain and let $N : R \to \mathbb{Z}^+ \cup \{0\}$ be a norm on R. The ring R is Euclidean with respect to N if for any $a, b \in R$ with $b \neq 0$, there exist elements q and r in R with

$$a = qb + r \qquad \text{with } r = 0 \text{ or } N(r) < N(b)\,.$$

Suppose now that this condition is weakened, namely that for any $a, b \in R$ with $b \neq 0$, there exist elements q, q' and r, r' in R with

$$a = qb + r, \quad b = q'r + r' \qquad \text{with } r' = 0 \text{ or } N(r') < N(b)\,,$$

i.e. the remainder after two divisions is smaller. Call such a domain a *2-stage Euclidean Domain*.

(a) Prove that iterating the divisions in a 2-stage Euclidean Domain produces a greatest common divisor of a and b which is a linear combination of a and b. Conclude that every *finitely generated* ideal of a 2-stage Euclidean Domain is principal. (There are 2-stage Euclidean domains that are *not* P.I.D.s, however.) [Imitate the proof of Theorem 4.]

(b) Prove that a 2-stage Euclidean Domain in which every nonzero nonunit can be factored into a finite number of irreducibles is a Unique Factorization Domain. [Prove first that irreducible elements are prime, as follows. If p is irreducible and $p|ab$ with p not dividing a, use part (a) to write $px + ay = 1$ for some x, y. Multiply through by b to conclude that $p|b$, so p is prime. Now follow the proof of uniqueness in Theorem 12.]

(c) Make the obvious generalization to define the notion of a k-stage Euclidean Domain for any integer $k \geq 1$. Prove that the statements in (a) and (b) remain valid if "2-stage Euclidean" is replaced by "k-stage Euclidean".

Remarks: There are examples of rings which are 2-stage Euclidean but are not Euclidean. There are also examples of rings which are not Euclidean with respect to a given norm but which are k-stage Euclidean with respect to the norm (for example, the ring $\mathbb{Z}[\sqrt{14}]$ is not Euclidean with respect to the usual norm $N(a + b\sqrt{14}) = |a^2 - 14b^2|$, but is 2-stage Euclidean with respect to this norm). The k-stage Euclidean condition is also related to the question of whether the group $GL_n(R)$ of invertible $n \times n$ matrices with entries from R is generated by elementary matrices (matrices with 1's along the main diagonal, a single 1 somewhere off the main diagonal, and 0's elsewhere).

9

Polynomial Rings

We begin this chapter on polynomial rings with a summary of facts from the preceding two chapters (with references where needed). The basic definitions were given in slightly greater detail in Section 7.2. For convenience, the ring R will always be a commutative ring with identity $1 \neq 0$.

9.1 DEFINITIONS AND BASIC PROPERTIES

The polynomial ring $R[x]$ in the indeterminate x with coefficients from R is the set of all formal sums $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ with $n \geq 0$ and each $a_i \in R$. If $a_n \neq 0$ then the polynomial is of degree n and $a_n x^n$ is the leading term. The polynomial is monic if $a_n = 1$. Addition of polynomials is "componentwise"

$$\sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} b_i x^i = \sum_{i=0}^{n} (a_i + b_i) x^i$$

(here a_n or b_n may be zero in order for addition of polynomials of different degrees to be defined). Multiplication is performed by first defining $(ax^i)(bx^j) = abx^{i+j}$ and then extending to all polynomials by the distributive laws so that in general

$$\left(\sum_{i=0}^{n} a_i x^i\right) \times \left(\sum_{i=0}^{m} b_i x^i\right) = \sum_{k=0}^{n+m} \left(\sum_{i=0}^{k} a_i b_{k-i}\right) x^k.$$

In this way $R[x]$ is a commutative ring with identity (the identity 1 from R) in which we identify R with the subring of constant polynomials.

We have already noted that if R is an integral domain then the leading term of a product of polynomials is the product of the leading terms of the factors. The following is Proposition 4 of Section 7.2 which we record here for completeness.

Proposition 1. Let R be an integral domain and let $p(x), q(x) \in R[x]$. Then

(1) degree $p(x)q(x)$ = degree $p(x)$ + degree $q(x)$

(2) the units of $R[x]$ are just the units of R

(3) $R[x]$ is an integral domain.

Recall also that if R is an integral domain, the quotient field of $R[x]$ consists of all quotients $\dfrac{p(x)}{q(x)}$ where $q(x)$ is not the zero polynomial (and is called the field of rational functions in x with coefficients in R).

The next result describes a relation between the ideals of R and those of $R[x]$.

Proposition 2. Let I be an ideal of the ring R and let $(I) = I[x]$ denote the ideal of $R[x]$ generated by I (the set of polynomials with coefficients in I). Then

$$R[x]/(I) \cong (R/I)[x].$$

In particular, if I is a prime ideal of R then (I) is a prime ideal of $R[x]$.

Proof: There is a natural map $\varphi : R[x] \to (R/I)[x]$ given by reducing each of the coefficients of a polynomial modulo I. The definition of addition and multiplication in these two rings shows that φ is a ring homomorphism. The kernel is precisely the set of polynomials each of whose coefficients is an element of I, so that $\ker \varphi = I[x] = (I)$, which proves the first part of the proposition. The last statement follows from Proposition 1, since if I is a prime ideal in R, then R/I is an integral domain, hence also $(R/I)[x]$ is an integral domain. This shows if I is a prime ideal of R, then (I) is a prime ideal of $R[x]$.

Note that it is not true that if I is a maximal ideal of R then (I) is a maximal ideal of $R[x]$. However, if I is maximal in R then the ideal of $R[x]$ generated by I and x is maximal in $R[x]$.

We now give an example of the "reduction homomorphism" of Proposition 2 which will be useful on a number of occasions later ("reduction homomorphisms" were also discussed at the end of Section 7.3 with reference to reducing the integers mod n) .

Example

Let $R = \mathbb{Z}$ and consider the ideal $n\mathbb{Z}$ of $\mathbb{Z}$. Then the isomorphism above can be written

$$\mathbb{Z}[x]/n\mathbb{Z}[x] \cong \mathbb{Z}/n\mathbb{Z}[x]$$

and the natural projection map of $\mathbb{Z}[x]$ to $\mathbb{Z}/n\mathbb{Z}[x]$ by reducing the coefficients modulo n is a ring homomorphism. If n is composite, then the quotient ring is not an integral domain. If, however, n is a prime p, then $\mathbb{Z}/p\mathbb{Z}$ is a field and so $\mathbb{Z}/p\mathbb{Z}[x]$ is an integral domain (in fact, a Euclidean Domain, as we shall see shortly). We also see that the set of polynomials whose coefficients are divisible by p is a prime ideal in $\mathbb{Z}[x]$.

We close this section with a description of the natural extension to polynomial rings in *several* variables.

Definition. The *polynomial ring in the variables* $x_1, x_2, \ldots, x_n$ *with coefficients in* R, denoted $R[x_1, x_2, \ldots, x_n]$, is defined inductively by

$$R[x_1, x_2, \ldots, x_n] = R[x_1, x_2, \ldots, x_{n-1}][x_n]$$

This definition means that we can consider polynomials in n variables with coefficients in R simply as polynomials in *one* variable (say x_n) but now with coefficients that are themselves *polynomials in* $n-1$ *variables*. In a slightly more concrete formulation, a polynomial in $x_1, x_2, \ldots, x_n$ with coefficients in R is a finite sum of *monomial* terms, i.e. a finite sum of terms of the form

$$ax_1^{d_1}x_2^{d_2}\ldots x_n^{d_n}$$

where $a \in R$ and the d_i are nonnegative integers. The exponent d_i is called the *degree in* x_i of the monomial and the sum

$$d = d_1 + d_2 + \cdots + d_n$$

is called the *degree* of the monomial. The ordered n-tuple $(d_1, d_2, \ldots, d_n)$ is the *multidegree* of the monomial. The *degree* of a polynomial is the largest degree of any of its monomial terms. The ring $R[x_1, x_2, \ldots, x_n]$ is again a commutative ring with identity.

Finally, to define a polynomial ring in an *arbitrary* number of variables with coefficients in R we take finite sums of monomials of the type above (but where the variables are not restricted to just $x_1, \ldots, x_n$), with the natural addition and multiplication. Alternatively, we could define this ring as the *union* of *all* the polynomial rings in a *finite* number of the variables being considered.

Example

The polynomial ring $\mathbb{Z}[x, y]$ in two variables x and y with integer coefficients consists of all finite sums of monomial terms of the form ax^iy^j (of degree $i+j$). For example,

$$P(x, y) = 2x^3 + xy - y^2$$

and

$$Q(x, y) = -3xy + 2y^2 + x^2y^3$$

are both elements of $\mathbb{Z}[x, y]$, of degrees 3 and 5, respectively. We have

$$P(x, y) + Q(x, y) = 2x^3 - 2xy + y^2 + x^2y^3$$

and

$$P(x, y)Q(x, y) = -6x^4y + 4x^3y^2 + 2x^5y^3 - 3x^2y^2 + 5xy^3 + x^3y^4 - 2y^4 - x^2y^5,$$

a polynomial of degree 8. To view this last polynomial, say, as a polynomial in y with coefficients in $\mathbb{Z}[x]$ as in the definition of several variable polynomial rings above, we would write the polynomial in the form

$$(-6x^4)y + (4x^3 - 3x^2)y^2 + (2x^5 + 5x)y^3 + (x^3 - 2)y^4 - (x^2)y^5.$$

Each of the statements in Proposition 1 is true for polynomial rings with an arbitrary number of variables. This follows by induction for finitely many variables and from the definition in terms of unions in the case of polynomial rings in arbitrarily many variables.

EXERCISES

1. Let $P(x, y, z) = 2x^2y - 3xy^3z + 4y^2z^5$ and $Q(x, y, z) = 7x^2 + 5x^2y^3z^4 - 3x^2z^3$ be polynomials in $\mathbb{Z}[x, y, z]$.
   (a) Write each of P and Q as a polynomial in x with coefficients in $\mathbb{Z}[y, z]$.
   (b) Find the degree of each of P and Q.
   (c) Find the degree of P and Q in each of the three variables x, y and z.
   (d) Compute PQ and find the degree of PQ in each of the three variables x, y and z.
   (e) Write PQ as a polynomial in the variable z with coefficients in $\mathbb{Z}[x, y]$.
2. Repeat the preceding exercise under the assumption that the coefficients of P and Q are in $\mathbb{Z}/3\mathbb{Z}$.
3. If R is a commutative ring and $x_1, x_2, \ldots, x_n$ are independent variables over R, prove that $R[x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}]$ is isomorphic to $R[x_1, x_2, \ldots, x_n]$ for any permutation π of $\{1, 2, \ldots, n\}$.
4. Prove that the ideals (x) and (x, y) are prime ideals in $\mathbb{Q}[x, y]$ but only the latter ideal is a maximal ideal.
5. Prove that (x, y) and $(2, x, y)$ are prime ideals in $\mathbb{Z}[x, y]$ but only the latter ideal is a maximal ideal.
6. Prove that (x, y) is not a principal ideal in $\mathbb{Q}[x, y]$.
7. Let R be a commutative ring with 1. Prove that a polynomial ring in more than one variable over R is not a Principal Ideal Domain.
8. Let R be a Noetherian ring (cf. Exercise 8 in Section 8.3). Prove that the polynomial ring $R[x_1, x_2, \ldots, x_n]$ is a Noetherian ring. [Use Exercise 11, Section 8.3.]
9. Prove that a polynomial ring in infinitely many variables with coefficients in any commutative ring is not a Noetherian ring.
10. Prove that the ring $\mathbb{Z}[x_1, x_2, x_3, \ldots]/(x_1x_2, x_3x_4, x_5x_6, \ldots)$ contains infinitely many minimal prime ideals (cf. Exercise 39 of Section 7.4).
11. Show that the radical of the ideal $I = (x, y^2)$ in $\mathbb{Q}[x, y]$ is (x, y) (cf. Exercise 30, Section 7.4). Deduce that I is a primary ideal that is not a power of a prime ideal (cf. Exercise 44, Section 7.4).
12. Let $R = \mathbb{Q}[x, y, z]$ and let bars denote passage to $\mathbb{Q}[x, y, z]/(xy - z^2)$. Prove that $\overline{P} = (\overline{x}, \overline{z})$ is a prime ideal. Show that $\overline{xy} \in \overline{P}^2$ but that no power of $\overline{y}$ lies in $\overline{P}^2$. (This shows $\overline{P}$ is a prime ideal whose square is *not* a primary ideal — cf. Exercise 44, Section 7.4).
13. Prove that the rings $F[x, y]/(y^2 - x)$ and $F[x, y]/(y^2 - x^2)$ are not isomorphic for any field F.

9.2 POLYNOMIAL RINGS OVER FIELDS I

We now consider more carefully the situation where the coefficient ring is a *field* F. We can define a *norm* on $F[x]$ by defining $N(P(x)) =$ degree of $P(x)$ (where we set $N(0) = 0$). From elementary algebra we know that we can divide one polynomial with, say, rational coefficients by another (nonzero) polynomial with rational coefficients to obtain a quotient and remainder. The same is true over any field.

Theorem 3. Let F be a field. The polynomial ring $F[x]$ is a Euclidean Domain. Specifically, if $a(x)$ and $b(x)$ are two polynomials in $F[x]$ with $b(x)$ nonzero, then there

are *unique* $q(x)$ and $r(x)$ in $F[x]$ such that

$$a(x) = q(x)b(x) + r(x) \qquad \text{with } r(x) = 0 \text{ or degree } r(x) < \text{degree } b(x)\,.$$

Proof: If $a(x)$ is the zero polynomial then take $q(x) = r(x) = 0$. We may therefore assume $a(x) \neq 0$ and prove the existence of $q(x)$ and $r(x)$ by induction on $n = \text{degree } a(x)$. Let $b(x)$ have degree m. If $n < m$ take $q(x) = 0$ and $r(x) = a(x)$. Otherwise $n \geq m$. Write

$$a(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

and

$$b(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0.$$

Then the polynomial $a'(x) = a(x) - \dfrac{a_n}{b_m}x^{n-m}b(x)$ is of degree less than n (we have arranged to subtract the leading term from $a(x)$). Note that this polynomial is well defined because the coefficients are taken from a *field* and $b_m \neq 0$. By induction then, there exist polynomials $q'(x)$ and $r(x)$ with

$$a'(x) = q'(x)b(x) + r(x) \qquad \text{with } r(x) = 0 \text{ or degree } r(x) < \text{degree } b(x)\,.$$

Then, letting $q(x) = q'(x) + \dfrac{a_n}{b_m}x^{n-m}$ we have

$$a(x) = q(x)b(x) + r(x) \qquad \text{with } r(x) = 0 \text{ or degree } r(x) < \text{degree } b(x)$$

completing the induction step.

As for the uniqueness, suppose $q_1(x)$ and $r_1(x)$ also satisfied the conditions of the theorem. Then both $a(x) - q(x)b(x)$ and $a(x) - q_1(x)b(x)$ are of degree less than $m = \text{degree } b(x)$. The difference of these two polynomials, i.e. $b(x)(q(x) - q_1(x))$ is also of degree less than m. But the degree of the product of two nonzero polynomials is the sum of their degrees (since F is an integral domain), hence $q(x) - q_1(x)$ must be 0, that is, $q(x) = q_1(x)$. This implies $r(x) = r_1(x)$, completing the proof.

Corollary 4. If F is a field, then $F[x]$ is a Principal Ideal Domain and a Unique Factorization Domain.

Proof: This is immediate from the results of the last chapter.

Recall also from Corollary 7 in Section 8.2 that if R is any commutative ring such that $R[x]$ is a Principal Ideal Domain (or Euclidean Domain) then R must be a field. We shall see in the next section, however, that $R[x]$ is a Unique Factorization Domain whenever R itself is a Unique Factorization Domain.

Examples

(1) By the above remarks the ring $\mathbb{Z}[x]$ is not a Principal Ideal Domain. As we have already seen (Example 3 beginning of Section 7.4) the ideal $(2, x)$ is not principal in this ring.

(2) $\mathbb{Q}[x]$ is a Principal Ideal Domain since the coefficients lie in the field $\mathbb{Q}$. The ideal generated in $\mathbb{Z}[x]$ by 2 and x is not principal in the subring $\mathbb{Z}[x]$ of $\mathbb{Q}[x]$. However, the ideal generated in $\mathbb{Q}[x]$ is principal; in fact it is the entire ring (so has 1 as a generator) since 2 is a unit in $\mathbb{Q}[x]$.

(3) If p is a prime, the ring $\mathbb{Z}/p\mathbb{Z}[x]$ obtained by reducing $\mathbb{Z}[x]$ modulo the prime ideal (p) is a Principal Ideal Domain, since the coefficients lie in the field $\mathbb{Z}/p\mathbb{Z}$. This example shows that the quotient of a ring which is not a Principal Ideal Domain *may* be a Principal Ideal Domain. To follow the ideal $(2, x)$ above in this example, note that if $p = 2$, then the ideal $(2, x)$ reduces to the ideal (x) in the quotient $\mathbb{Z}/2\mathbb{Z}[x]$, which is a proper (maximal) ideal. If $p \neq 2$, then 2 is a unit in the quotient, so the ideal $(2, x)$ reduces to the entire ring $\mathbb{Z}/p\mathbb{Z}[x]$.

(4) $\mathbb{Q}[x, y]$, the ring of polynomials in two variables with rational coefficients, is *not* a Principal Ideal Domain since this ring is $\mathbb{Q}[x][y]$ and $\mathbb{Q}[x]$ is not a field (any element of positive degree is not invertible). It is an exercise to see that the ideal (x, y) is not a principal ideal in this ring. We shall see shortly that $\mathbb{Q}[x, y]$ *is* a Unique Factorization Domain.

We note that the quotient and remainder in the Division Algorithm applied to $a(x), b(x) \in F[x]$ are *independent of field extensions* in the following sense. Suppose the field F is contained in the field E and $a(x) = Q(x)b(x) + R(x)$ for some $Q(x)$, $R(x)$ satisfying the conditions of Theorem 3 in $E[x]$. Write $a(x) = q(x)b(x) + r(x)$ for some $q(x), r(x) \in F[x]$ and apply the uniqueness condition of Theorem 3 in the ring $E[x]$ to deduce that $Q(x) = q(x)$ and $R(x) = r(x)$. In particular, $b(x)$ divides $a(x)$ in the ring $E[x]$ if and only if $b(x)$ divides $a(x)$ in $F[x]$. Also, the greatest common divisor of $a(x)$ and $b(x)$ (which can be obtained from the Euclidean Algorithm) is the same, once we make it unique by specifying it to be monic, whether these elements are viewed in $F[x]$ or in $E[x]$.

EXERCISES

Let F be a field and let x be an indeterminate over F.

1. Let $f(x) \in F[x]$ be a polynomial of degree $n \geq 1$ and let bars denote passage to the quotient $F[x]/(f(x))$. Prove that for each $\overline{g(x)}$ there is a unique polynomial $g_0(x)$ of degree $\leq n - 1$ such that $\overline{g(x)} = \overline{g_0(x)}$ (equivalently, the elements $\overline{1}, \overline{x}, \ldots, \overline{x^{n-1}}$ are a *basis* of the vector space $F[x]/(f(x))$ over F—in particular, the dimension of this space is n). [Use the Division Algorithm.]
2. Let F be a finite field of order q and let $f(x)$ be a polynomial in $F[x]$ of degree $n \geq 1$. Prove that $F[x]/(f(x))$ has q^n elements. [Use the preceding exercise.]
3. Let $f(x)$ be a polynomial in $F[x]$. Prove that $F[x]/(f(x))$ is a field if and only if $f(x)$ is irreducible. [Use Proposition 6, Section 8.2.]
4. Let F be a finite field. Prove that $F[x]$ contains infinitely many primes. (Note that over an infinite field the polynomials of degree 1 are an infinite set of primes in the ring of polynomials).
5. Exhibit *all* the ideals in the ring $F[x]/(p(x))$, where F is a field and $p(x)$ is a polynomial in $F[x]$ (describe them in terms of the factorization of $p(x)$).
6. Describe (briefly) the ring structure of the following rings:
   (a) $\mathbb{Z}[x]/(2)$, (b) $\mathbb{Z}[x]/(x)$, (c) $\mathbb{Z}[x]/(x^2)$, (d) $\mathbb{Z}[x, y]/(x^2, y^2, 2)$.
   Show that $\alpha^2 = 0$ or 1 for every α in the last ring and determine those elements with $\alpha^2 = 0$. Determine the characteristics of each of these rings (cf. Exercise 26, Section 7.3).
7. Determine all the ideals of the ring $\mathbb{Z}[x]/(2, x^3 + 1)$.
8. Determine the greatest common divisor of $a(x) = x^3 - 2$ and $b(x) = x + 1$ in $\mathbb{Q}[x]$ and write it as a linear combination (in $\mathbb{Q}[x]$) of $a(x)$ and $b(x)$.

9. Determine the greatest common divisor of $a(x) = x^5 + 2x^3 + x^2 + x + 1$ and $b(x) = x^5 + x^4 + 2x^3 + 2x^2 + 2x + 1$ in $\mathbb{Q}[x]$ and write it as a linear combination (in $\mathbb{Q}[x]$) of $a(x)$ and $b(x)$.

10. Determine the greatest common divisor of $a(x) = x^3 + 4x^2 + x - 6$ and $b(x) = x^5 - 6x + 5$ in $\mathbb{Q}[x]$ and write it as a linear combination (in $\mathbb{Q}[x]$) of $a(x)$ and $b(x)$.

11. This exercise introduces a noncommutative ring which is a "right" Euclidean Domain (and a "left" Principal Ideal Domain) but is not a "left" Euclidean Domain (and not a "right" Principal Ideal Domain). Let F be a field of characteristic p in which not every element is a pth power: $F \neq F^p$ (for example the field $F = \mathbb{F}_p(t)$ of rational functions in the variable t with coefficients in $\mathbb{F}_p$ is such a field). Let $R = F\{x\}$ be the "twisted" polynomial ring of polynomials $\sum_{i=0}^{n} a_i x^i$ in x with coefficients in F with the usual (termwise) addition

$$\sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} b_i x^i = \sum_{i=0}^{n} (a_i + b_i) x^i$$

but with a noncommutative multiplication defined by

$$\left(\sum_{i=0}^{n} a_i x^i\right)\left(\sum_{j=0}^{m} b_j x^j\right) = \sum_{k=0}^{n+m}\left(\sum_{i+j=k} a_i b_j^{p^i}\right) x^k .$$

This multiplication arises from defining $xa = a^p x$ for every $a \in F$ (so the powers of x do not commute with the coefficients) and extending in a natural way. Let N be the norm defined by taking the degree of a polynomial in R: $N(f) = \deg(f)$.

(a) Show that $x^k a = a^{p^k} x^k$ for every $a \in F$ and every integer $k \geq 0$ and that R is a ring with this definition of multiplication. [Use the fact that $(a+b)^p = a^p + b^p$ for every $a, b \in F$ since F has characteristic p, so also $(a+b)^{p^k} = a^{p^k} + b^{p^k}$ for every $a, b \in F$.]

(b) Prove that the degree of a product of two elements of R is the sum of the degrees of the elements. Prove that R has no zero divisors.

(c) Prove that R is "right Euclidean" with respect to N, i.e. for any polynomials $f, g \in R$ with $g \neq 0$, there exist polynomials q and r in R with

$$f = qg + r \qquad \text{with } r = 0 \text{ or } \deg(r) < \deg(g) .$$

Use this to prove that every *left* ideal of R is principal.

(d) Let $f = \theta x$ for some $\theta \in F$, $\theta \notin F^p$ and let $g = x$. Prove that there are no polynomials q and r in R with

$$f = gq + r \qquad \text{with } r = 0 \text{ or } \deg(r) < \deg(g) ,$$

so in particular R is not "left Euclidean" with respect to N. Prove that the right ideal of R generated by x and θx is not principal. Conclude that R is not "left Euclidean" with respect to *any* norm.

9.3 POLYNOMIAL RINGS THAT ARE UNIQUE FACTORIZATION DOMAINS

We have seen in Proposition 1 that if R is an integral domain then $R[x]$ is also an integral domain. Also, such an R can be embedded in its field of fractions F (Theorem 15, Section 7.5), so that $R[x] \subseteq F[x]$ as a subring, and $F[x]$ is a Euclidean Domain

(hence a Principal Ideal Domain and a Unique Factorization Domain). Many computations for $R[x]$ may be accomplished in $F[x]$ at the expense of allowing fractional coefficients. This raises the immediate question of how computations (such as factorizations of polynomials) in $F[x]$ can be used to give information in $R[x]$.

For instance, suppose $p(x)$ is a polynomial in $R[x]$. Since $F[x]$ is a Unique Factorization Domain we can factor $p(x)$ uniquely into a product of irreducibles in $F[x]$. It is natural to ask whether we can do the same in $R[x]$, i.e. is $R[x]$ a Unique Factorization Domain? In general the answer is no because if $R[x]$ were a Unique Factorization Domain, the constant polynomials would have to be uniquely factored into irreducible elements of $R[x]$, necessarily of degree 0 since the degrees of products add, that is, R would itself have to be a Unique Factorization Domain. Thus if R is an integral domain which is not a Unique Factorization Domain, $R[x]$ cannot be a Unique Factorization Domain. On the other hand, it turns out that if R is a Unique Factorization Domain, then $R[x]$ is also a Unique Factorization Domain. The method of proving this is to first factor uniquely in $F[x]$ and then "clear denominators" to obtain a unique factorization in $R[x]$. The first step in making this precise is to compare the factorization of a polynomial in $F[x]$ to a factorization in $R[x]$.

Proposition 5. (Gauss' Lemma) Let R be a Unique Factorization Domain with field of fractions F and let $p(x) \in R[x]$. If $p(x)$ is reducible in $F[x]$ then $p(x)$ is reducible in $R[x]$. More precisely, if $p(x) = A(x)B(x)$ for some nonconstant polynomials $A(x), B(x) \in F[x]$, then there are nonzero elements $r, s \in F$ such that $rA(x) = a(x)$ and $sB(x) = b(x)$ both lie in $R[x]$ and $p(x) = a(x)b(x)$ is a factorization in $R[x]$.

Proof: The coefficients of the polynomials on the right in the equation $p(x) = A(x)B(x)$ are elements in the field F, hence are quotients of elements from the Unique Factorization Domain R. Multiplying through by a common denominator for all these coefficients, we obtain an equation $dp(x) = a'(x)b'(x)$ where now $a'(x)$ and $b'(x)$ are elements of $R[x]$ and d is a nonzero element of R. If d is a unit in R, the proposition is true with $a(x) = d^{-1}a'(x)$ and $b(x) = b'(x)$. Assume d is not a unit and write d as a product of irreducibles in R, say $d = p_1 \dots p_n$. Since p_1 is irreducible in R, the ideal (p_1) is prime (cf. Proposition 10, Section 8.3), so by Proposition 2 above, the ideal $p_1R[x]$ is prime in $R[x]$ and $(R/p_1R)[x]$ is an integral domain. Reducing the equation $dp(x) = a'(x)b'(x)$ modulo p_1, we obtain the equation $0 = \overline{a'(x)}\,\overline{b'(x)}$ in this integral domain (the bars denote the images of these polynomials in the quotient ring), hence one of the two factors, say $\overline{a'(x)}$ must be 0. But this means all the coefficients of $a'(x)$ are divisible by p_1, so that $\frac{1}{p_1}a'(x)$ also has coefficients in R. In other words, in the equation $dp(x) = a'(x)b'(x)$ we can cancel a factor of p_1 from d (on the left) and from either $a'(x)$ or $b'(x)$ (on the right) and still have an equation in $R[x]$. But now the factor d on the left hand side has one fewer irreducible factors. Proceeding in the same fashion with each of the remaining factors of d, we can cancel all of the factors of d into the two polynomials on the right hand side, leaving an equation $p(x) = a(x)b(x)$ with $a(x), b(x) \in R[x]$ and with $a(x), b(x)$ being F-multiples of $A(x), B(x)$, respectively. This completes the proof.

Note that we cannot prove that $a(x)$ and $b(x)$ are necessarily R-multiples of $A(x), B(x)$, respectively, because, for example, we could factor x^2 in $\mathbb{Q}[x]$ with $A(x) = 2x$ and $B(x) = \frac{1}{2}x$ but no *integer* multiples of $A(x)$ and $B(x)$ give a factorization of x^2 in $\mathbb{Z}[x]$.

The elements of the ring R become *units* in the Unique Factorization Domain $F[x]$ (the units in $F[x]$ being the nonzero elements of F). For example, $7x$ factors in $\mathbb{Z}[x]$ into a product of two irreducibles: 7 and x (so $7x$ is not irreducible in $\mathbb{Z}[x]$), whereas $7x$ is the unit 7 times the irreducible x in $\mathbb{Q}[x]$ (so $7x$ is irreducible in $\mathbb{Q}[x]$). The following corollary shows that this is essentially the *only* difference between the irreducible elements in $R[x]$ and those in $F[x]$.

Corollary 6. Let R be a Unique Factorization Domain, let F be its field of fractions and let $p(x) \in R[x]$. Suppose the greatest common divisor of the coefficients of $p(x)$ is 1. Then $p(x)$ is irreducible in $R[x]$ if and only if it is irreducible in $F[x]$. In particular, if $p(x)$ is a monic polynomial that is irreducible in $R[x]$, then $p(x)$ is irreducible in $F[x]$.

Proof: By Gauss' Lemma above, if $p(x)$ is reducible in $F[x]$, then it is reducible in $R[x]$. Conversely, the assumption on the greatest common divisor of the coefficients of $p(x)$ implies that if it is reducible in $R[x]$, then $p(x) = a(x)b(x)$ where neither $a(x)$ nor $b(x)$ are constant polynomials in $R[x]$. This same factorization shows that $p(x)$ is reducible in $F[x]$, completing the proof.

Theorem 7. R is a Unique Factorization Domain if and only if $R[x]$ is a Unique Factorization Domain.

Proof: We have indicated above that $R[x]$ a Unique Factorization Domain forces R to be a Unique Factorization Domain. Suppose conversely that R is a Unique Factorization Domain, F is its field of fractions and $p(x)$ is a nonzero element of $R[x]$. Let d be the greatest common divisor of the coefficients of $p(x)$, so that $p(x) = dp'(x)$, where the g.c.d. of the coefficients of $p'(x)$ is 1. Such a factorization of $p(x)$ is unique up to a change in d (so up to a unit in R), and since d can be factored uniquely into irreducibles in R (and these are also irreducibles in the larger ring $R[x]$), it suffices to prove that $p'(x)$ can be factored uniquely into irreducibles in $R[x]$. Thus we may assume that the greatest common divisor of the coefficients of $p(x)$ is 1. We may further assume $p(x)$ is not a unit in $R[x]$, i.e. degree $p(x) > 0$.

Since $F[x]$ is a Unique Factorization Domain, $p(x)$ can be factored uniquely into irreducibles in $F[x]$. By Gauss' Lemma, such a factorization implies there is a factorization of $p(x)$ in $R[x]$ whose factors are F-multiples of the factors in $F[x]$. Since the greatest common divisor of the coefficients of $p(x)$ is 1, the g.c.d. of the coefficients in each of these factors in $R[x]$ must be 1. By Corollary 6, each of these factors is an irreducible in $R[x]$. This shows that $p(x)$ can be written as a finite product of irreducibles in $R[x]$.

The uniqueness of the factorization of $p(x)$ follows from the uniqueness in $F[x]$. Suppose

$$p(x) = q_1(x) \cdots q_r(x) = q_1'(x) \cdots q_s'(x)$$

are two factorizations of $p(x)$ into irreducibles in $R[x]$. Since the g.c.d. of the coefficients of $p(x)$ is 1, the same is true for each of the irreducible factors above — in particular, each has positive degree. By Corollary 6, each $q_i(x)$ and $q'_j(x)$ is an irreducible in $F[x]$. By unique factorization in $F[x]$, $r = s$ and, possibly after rearrangement, $q_i(x)$ and $q'_i(x)$ are associates in $F[x]$ for all $i \in \{1, \ldots, r\}$. It remains to show they are associates in $R[x]$. Since the units of $F[x]$ are precisely the elements of $F^\times$ we need to consider when $q(x) = \frac{a}{b}q'(x)$ for some $q(x), q'(x) \in R[x]$ and nonzero elements a, b of R, where the greatest common divisor of the coefficients of each of $q(x)$ and $q'(x)$ is 1. In this case $bq(x) = aq'(x)$; the g.c.d. of the coefficients on the left hand side is b and on the right hand side is a. Since in a Unique Factorization Domain the g.c.d. of the coefficients of a nonzero polynomial is unique up to units, $a = ub$ for some unit u in R. Thus $q(x) = uq'(x)$ and so $q(x)$ and $q'(x)$ are associates in R as well. This completes the proof.

Corollary 8. If R is a Unique Factorization Domain, then a polynomial ring in an arbitrary number of variables with coefficients in R is also a Unique Factorization Domain.

Proof: For finitely many variables, this follows by induction from Theorem 7, since a polynomial ring in n variables can be considered as a polynomial ring in one variable with coefficients in a polynomial ring in $n - 1$ variables. The general case follows from the definition of a polynomial ring in an arbitrary number of variables as the union of polynomial rings in finitely many variables.

Examples

(1) $\mathbb{Z}[x]$, $\mathbb{Z}[x, y]$, etc. are Unique Factorization Domains. The ring $\mathbb{Z}[x]$ gives an example of a Unique Factorization Domain that is not a Principal Ideal Domain.
(2) Similarly, $\mathbb{Q}[x]$, $\mathbb{Q}[x, y]$, etc. are Unique Factorization Domains.

We saw earlier that if R is a Unique Factorization Domain with field of fractions F and $p(x) \in R[x]$, then we can factor out the greatest common divisor d of the coefficients of $p(x)$ to obtain $p(x) = dp'(x)$, where $p'(x)$ is irreducible in both $R[x]$ and $F[x]$. Suppose now that R is an *arbitrary* integral domain with field of fractions F. In R the notion of greatest common divisor may not make sense, however one might still ask if, say, a *monic* polynomial which is irreducible in $R[x]$ is still irreducible in $F[x]$ (i.e. whether the last statement in Corollary 6 is true).

Note first that if a monic polynomial $p(x)$ is reducible, it must have a factorization $p(x) = a(x)b(x)$ in $R[x]$ with both $a(x)$ and $b(x)$ *monic, nonconstant* polynomials (recall that the leading term of $p(x)$ is the product of the leading terms of the factors, so the leading coefficients of both $a(x)$ and $b(x)$ are units — we can thus arrange these to be 1). In other words, a nonconstant *monic* polynomial $p(x)$ is irreducible if and only if it cannot be factored as a product of two *monic* polynomials of smaller degree.

We now see that it is not true that if R is an arbitrary integral domain and $p(x)$ is a monic irreducible polynomial in $R[x]$, then $p(x)$ is irreducible in $F[x]$. For example, let $R = \mathbb{Z}[2i] = \{a + 2bi \mid a, b \in \mathbb{Z}\}$ (a subring of the complex numbers) and let $p(x) = x^2 + 1$. Then the fraction field of R is $F = \{a + bi \mid a, b \in \mathbb{Q}\}$. The polynomial $p(x)$ factors uniquely into a product of two linear factors in $F[x]$: $x^2 + 1 = (x - i)(x + i)$

so in particular, $p(x)$ *is reducible in* $F[x]$. Neither of these factors lies in $R[x]$ (because $i \notin R$) so $p(x)$ *is irreducible in* $R[x]$. In particular, by Corollary 6, $\mathbb{Z}[2i]$ *is not a Unique Factorization Domain.*

EXERCISES

1. Let R be an integral domain with quotient field F and let $p(x)$ be a monic polynomial in $R[x]$. Assume that $p(x) = a(x)b(x)$ where $a(x)$ and $b(x)$ are monic polynomials in $F[x]$ of smaller degree than $p(x)$. Prove that if $a(x) \notin R[x]$ then R is not a Unique Factorization Domain. Deduce that $\mathbb{Z}[2\sqrt{2}]$ is not a U.F.D.
2. Prove that if $f(x)$ and $g(x)$ are polynomials with rational coefficients whose product $f(x)g(x)$ has integer coefficients, then the product of any coefficient of $g(x)$ with any coefficient of $f(x)$ is an integer.

9.4 IRREDUCIBILITY CRITERIA

If R is a Unique Factorization Domain, then by Corollary 8 a polynomial ring in any number of variables with coefficients in R is also a Unique Factorization Domain. It is of interest then to determine the irreducible elements in such a polynomial ring, particularly in the ring $R[x]$. In the one–variable case, a nonconstant monic polynomial is irreducible in $R[x]$ if it cannot be factored as the product of two other polynomials of smaller degrees. Determining whether a polynomial has factors is frequently difficult to check, particularly for polynomials of large degree in several variables. The purpose of irreducibility criteria is to give an easier mechanism for determining when some types of polynomials are irreducible.

For the most part we restrict attention to polynomials in one variable where the coefficient ring is a Unique Factorization Domain. By Gauss' Lemma it suffices to consider factorizations in $F[x]$ where F is the field of fractions of R (although we shall occasionally consider questions of irreducibility when the coefficient ring is just an integral domain). The next proposition considers when there is a factor of degree one (a *linear* factor).

Proposition 9. Let F be a field and let $p(x) \in F[x]$. Then $p(x)$ has a factor of degree one if and only if $p(x)$ has a root in F, i.e. there is an $\alpha \in F$ with $p(\alpha) = 0$.

Proof: If $p(x)$ has a factor of degree one, then since F is a field, we may assume the factor is monic, i.e. is of the form $(x - \alpha)$ for some $\alpha \in F$. But then $p(\alpha) = 0$. Conversely, suppose $p(\alpha) = 0$. By the Division Algorithm in $F[x]$ we may write

$$p(x) = q(x)(x - \alpha) + r$$

where r is a constant. Since $p(\alpha) = 0$, r must be 0, hence $p(x)$ has $(x - \alpha)$ as a factor.

Proposition 9 gives a criterion for irreducibility for polynomials of small degree:

Proposition 10. A polynomial of degree two or three over a field F is reducible if and only if it has a root in F.

Proof: This follows immediately from the previous proposition, since a polynomial of degree two or three is reducible if and only if it has at least one linear factor.

The next result limits the possibilities for roots of polynomials with integer coefficients (it is stated for $\mathbb{Z}[x]$ for convenience although it clearly generalizes to $R[x]$, where R is any Unique Factorization Domain).

Proposition 11. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a polynomial of degree n with integer coefficients. If $\frac{r}{s} \in \mathbb{Q}$ is in lowest terms (i.e. r and s are relatively prime integers) and $\frac{r}{s}$ is a root of $p(x)$, then r divides the constant term and s divides the leading coefficient of $p(x)$: $r \mid a_0$ and $s \mid a_n$. In particular, if $p(x)$ is a *monic* polynomial with integer coefficients and $p(d) \neq 0$ for all integers d dividing the constant term of $p(x)$, then $p(x)$ has no roots in $\mathbb{Q}$.

Proof: By hypothesis, $p(\frac{r}{s}) = 0 = a_n(\frac{r}{s})^n + a_{n-1}(\frac{r}{s})^{n-1} + \cdots + a_0$. Multiplying through by s^n gives

$$0 = a_n r^n + a_{n-1}r^{n-1}s + \cdots + a_0 s^n.$$

Thus $a_n r^n = s(-a_{n-1}r^{n-1} - \cdots - a_0 s^{n-1})$, so s divides $a_n r^n$. By assumption, s is relatively prime to r and it follows that $s \mid a_n$. Similarly, $a_0 s^n = r(-a_n r^{n-1} - \cdots - a_1 s^{n-1})$, so $r \mid a_0$. The last assertion of the proposition follows from the previous ones.

Examples

(1) The polynomial $x^3 - 3x - 1$ is irreducible in $\mathbb{Z}[x]$. To prove this, by Gauss' Lemma and Proposition 10 it suffices to show it has no rational roots. By Proposition 11 the only candidates for rational roots are integers which divide the constant term 1, namely ± 1. Substituting both 1 and -1 into the polynomial shows that these are not roots.

(2) For p any prime the polynomials $x^2 - p$ and $x^3 - p$ are irreducible in $\mathbb{Q}[x]$. This is because they have degrees ≤ 3 so it suffices to show they have no rational roots. By Proposition 11 the only candidates for roots are ± 1 and $\pm p$, but none of these give 0 when they are substituted into the polynomial.

(3) The polynomial $x^2 + 1$ is reducible in $\mathbb{Z}/2\mathbb{Z}[x]$ since it has 1 as a root, and it factors as $(x+1)^2$.

(4) The polynomial $x^2 + x + 1$ is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$ since it does not have a root in $\mathbb{Z}/2\mathbb{Z}$: $0^2 + 0 + 1 = 1$ and $1^2 + 1 + 1 = 1$.

(5) Similarly, the polynomial $x^3 + x + 1$ is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$.

This technique is limited to polynomials of low degree because it relies on the presence of a factor of degree one. A polynomial of degree 4, for example, may be the product of two irreducible quadratics, hence be reducible but have no linear factor. One fairly general technique for checking irreducibility uses Proposition 2 above and consists of reducing the coefficients modulo some ideal.

Proposition 12. Let I be a proper ideal in the integral domain R and let $p(x)$ be a nonconstant monic polynomial in $R[x]$. If the image of $p(x)$ in $(R/I)[x]$ cannot be factored in $(R/I)[x]$ into two polynomials of smaller degree, then $p(x)$ is irreducible in $R[x]$.

Proof: Suppose $p(x)$ cannot be factored in $(R/I)[x]$ but that $p(x)$ is reducible in $R[x]$. As noted at the end of the preceding section this means there are monic, nonconstant polynomials $a(x)$ and $b(x)$ in $R[x]$ such that $p(x) = a(x)b(x)$. By Proposition 2, reducing the coefficients modulo I gives a factorization in $(R/I)[x]$ with nonconstant factors, a contradiction.

This proposition indicates that if it is possible to find a proper ideal I such that the *reduced* polynomial cannot be factored, then the polynomial is itself irreducible. Unfortunately, there are examples of polynomials even in $\mathbb{Z}[x]$ which are irreducible but whose reductions modulo every ideal are reducible (so their irreducibility is not detectable by this technique). For example, the polynomial $x^4 + 1$ is irreducible in $\mathbb{Z}[x]$ but is reducible modulo every prime (we shall verify this in Chapter 14) and the polynomial $x^4 - 10x^2 + 17$ is irreducible in $\mathbb{Z}[x]$ but is reducible modulo every integer.

Examples

(1) Consider the polynomial $p(x) = x^2 + x + 1$ in $\mathbb{Z}[x]$. Reducing modulo 2, we see from Example 4 above that $p(x)$ is irreducible in $\mathbb{Z}[x]$. Similarly, $x^3 + x + 1$ is irreducible in $\mathbb{Z}[x]$ because it is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$.

(2) The polynomial $x^2 + 1$ is irreducible in $\mathbb{Z}[x]$ since it is irreducible in $\mathbb{Z}/3\mathbb{Z}[x]$ (no root in $\mathbb{Z}/3\mathbb{Z}$), but is reducible mod 2. This shows that the converse to Proposition 12 does not hold.

(3) The idea of reducing modulo an ideal to determine irreducibility can be used also in several variables, but some care must be exercised. For example, the polynomial $x^2 + xy + 1$ in $\mathbb{Z}[x, y]$ is irreducible since modulo the ideal (y) it is $x^2 + 1$ in $\mathbb{Z}[x]$, which is irreducible and of the same degree. In this sort of argument it is necessary to be careful about "collapsing." For example, the polynomial $xy + x + y + 1$ ($= (x + 1)(y + 1)$) is reducible, but appears irreducible modulo both (x) and (y). The reason for this is that nonunit polynomials in $\mathbb{Z}[x, y]$ can reduce to units in the quotient. To take account of this it is necessary to determine which elements in the original ring become units in the quotient. The elements in $\mathbb{Z}[x, y]$ which are units modulo (y), for example, are the polynomials in $\mathbb{Z}[x, y]$ with constant term ± 1 and all nonconstant terms divisible by y. The fact that $x^2 + xy + 1$ and its reduction mod (y) have the same degree therefore eliminates the possibility of a factor which is a unit modulo (y), but not a unit in $\mathbb{Z}[x, y]$ and gives the irreducibility of this polynomial.

A special case of this notion of reducing modulo an ideal to test for irreducibility which is frequently useful is *Eisenstein's Criterion*:

Proposition 13. (Eisenstein's Criterion) Let P be a prime ideal of the integral domain R and let $f(x) = x^n + a_{n-1} + \cdots + a_1 x + a_0$ be a polynomial in $R[x]$ (here $n \geq 1$). Suppose $a_{n-1}, \ldots, a_1, a_0$ are all elements of P and suppose a_0 is not an element of P^2. Then $f(x)$ is irreducible in $R[x]$.

Proof: Suppose $f(x)$ were reducible, say $f(x) = a(x)b(x)$ in $R[x]$, where $a(x)$ and $b(x)$ are nonconstant polynomials. Reducing this equation modulo P and using the assumptions on the coefficients of $f(x)$ we obtain the equation $x^n = \overline{a(x)}\,\overline{b(x)}$ in $(R/P)[x]$, where the bar denotes the polynomials with coefficients reduced mod P. Since P is a prime ideal, R/P is an integral domain, and it follows that both $\overline{a(x)}$ and $\overline{b(x)}$ have 0 constant term, i.e. the constant terms of both $a(x)$ and $b(x)$ are elements of P. But then the constant term a_0 of $f(x)$ as the product of these two would be an element of P^2, a contradiction.

Eisenstein's Criterion is most frequently applied to $\mathbb{Z}[x]$ so we state the result explicitly for this case:

Corollary 14. (Eisenstein's Criterion for $\mathbb{Z}[x]$) Let p be a prime in $\mathbb{Z}$ and let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in \mathbb{Z}[x]$, $n \geq 1$. Suppose p divides a_i for all $i \in \{0, 1, \ldots, n-1\}$ but that p^2 does not divide a_0. Then $f(x)$ is irreducible in both $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$.

Proof: This is simply a restatement of Proposition 13 in the case of the prime ideal (p) in $\mathbb{Z}$ together with Corollary 6.

Examples

(1) The polynomial $x^4 + 10x + 5$ in $\mathbb{Z}[x]$ is irreducible by Eisenstein's Criterion applied for the prime 5.

(2) If a is any integer which is divisible by some prime p but not divisible by p^2, then $x^n - a$ is irreducible in $\mathbb{Z}[x]$ by Eisenstein's Criterion. In particular, $x^n - p$ is irreducible for all positive integers n and so for $n \geq 2$ the nth roots of p are not rational numbers (i.e. this polynomial has no root in $\mathbb{Q}$).

(3) Consider the polynomial $f(x) = x^4 + 1$ mentioned previously. Eisenstein's Criterion does not apply directly to $f(x)$. The polynomial $g(x) = f(x+1)$ is $(x+1)^4 + 1 = x^4 + 4x^3 + 6x^2 + 4x + 2$, and Eisenstein's Criterion for the prime 2 shows that this polynomial is irreducible. It follows then that $f(x)$ must also be irreducible, since any factorization for $f(x)$ would provide a factorization for $g(x)$ (just replace x by $x+1$ in each of the factors). This example shows that Eisenstein's Criterion can sometimes be used to verify the irreducibility of a polynomial to which it does not immediately apply.

(4) As another example of this, let p be a prime and consider the polynomial

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1,$$

an example of a *cyclotomic polynomial* which we shall consider more thoroughly in Part IV. Again, Eisenstein's Criterion does not immediately apply, but it does apply for the prime p to the polynomial

$$\Phi_p(x+1) = \frac{(x+1)^p - 1}{x} = x^{p-1} + px^{p-2} + \cdots + \frac{p(p-1)}{2}x + p \in \mathbb{Z}[x]$$

since all the coefficients except the first are divisible by p by the Binomial Theorem. As before, this shows $\Phi_p(x)$ is irreducible in $\mathbb{Z}[x]$.

(5) As an example of the use of the more general Eisenstein's Criterion in Proposition 13 we mimic Example 2 above. Let $R = \mathbb{Q}[x]$ and let n be any positive integer. Consider the polynomial $X^n - x$ in the ring $R[X]$. The ideal (x) is prime in the coefficient ring R since $R/(x) = \mathbb{Q}[x]/(x)$ is the integral domain $\mathbb{Q}$. Eisenstein's Criterion for the ideal (x) of R applies directly to show that $X^n - x$ is irreducible in $R[X]$. Note that this construction works with $\mathbb{Q}$ replaced by any field or, indeed, by any integral domain.

There are now efficient algorithms for factoring polynomials over certain fields. For polynomials with integer coefficients these algorithms have been implemented in a number of computer packages and work quite rapidly for polynomials which have relatively small degree (≤ 20, say) and relatively small coefficients (approximately 10^{20}, say). An efficient algorithm for factoring polynomials over $\mathbb{F}_p$, called the Berlekamp Algorithm, is described in detail in the exercises at the end of Section 14.3.

EXERCISES

1. Determine whether the following polynomials are irreducible in the rings indicated. For those that are reducible, determine their factorization into irreducibles. The notation $\mathbb{F}_p$ denotes the finite field $\mathbb{Z}/p\mathbb{Z}$, p a prime.
   (a) $x^2 + x + 1$ in $\mathbb{F}_2[x]$.
   (b) $x^3 + x + 1$ in $\mathbb{F}_3[x]$.
   (c) $x^4 + 1$ in $\mathbb{F}_5[x]$.
   (d) $x^4 + 10x^2 + 1$ in $\mathbb{Z}[x]$.
2. Prove that the following polynomials are irreducible in $\mathbb{Z}[x]$:
   (a) $x^4 - 4x^3 + 6$
   (b) $x^6 + 30x^5 - 15x^3 + 6x - 120$
   (c) $x^4 + 4x^3 + 6x^2 + 2x + 1$ [Substitute $x - 1$ for x.]
   (d) $\dfrac{(x+2)^p - 2^p}{x}$, where p is an odd prime.
3. Show that the polynomial $(x-1)(x-2)\cdots(x-n) - 1$ is irreducible over $\mathbb{Z}$ for all $n \geq 1$.
4. Show that the polynomial $(x-1)(x-2)\cdots(x-n) + 1$ is irreducible over $\mathbb{Z}$ for all $n \geq 1$.
5. Find all the monic irreducible polynomials of degree ≤ 3 in $\mathbb{F}_2[x]$, and the same in $\mathbb{F}_3[x]$.
6. Construct fields of each of the following orders: (a) 9, (b) 49, (c) 8, (d) 81 (you may exhibit these as $F[x]/(f(x))$ for suitable F and f). [Use Exercises 2 and 3 in Section 2.]
7. Prove that $\mathbb{R}[x]/(x^2+1)$ is a field which is isomorphic to the complex numbers.
8. Prove that $K_1 = \mathbb{F}_{11}[x]/(x^2+1)$ and $K_2 = \mathbb{F}_{11}[y]/(y^2+2y+2)$ are both fields with 121 elements. Prove that the map which sends the element $p(\bar{x})$ of K_1 to the element $p(\bar{y}+1)$ of K_2 (where p is any polynomial with coefficients in $\mathbb{F}_{11}$) is well defined and gives a ring (hence field) isomorphism from K_1 to K_2.
9. Prove that the polynomial $x^2 - \sqrt{2}$ is irreducible over $\mathbb{Z}[\sqrt{2}]$ (you may use the fact that $\mathbb{Z}[\sqrt{2}]$ is a U.F.D. — cf. Exercise 5(b) of Section 8.3).
10. Prove that the polynomial $p(x) = x^4 - 4x^2 + 8x + 2$ is irreducible over the field $F = \mathbb{Q}(\sqrt{-2}) = \{a + b\sqrt{-2} \mid a, b \in \mathbb{Q}\}$. [First use the method of Proposition 11 for the Unique Factorization Domain $\mathbb{Z}[\sqrt{-2}]$ (cf. Exercise 8, Section 8.1) to show that if $\alpha \in \mathbb{Z}[\sqrt{-2}]$ is a root of $p(x)$ then α is a divisor of 2 in $\mathbb{Z}[\sqrt{-2}]$. Conclude that α must be ± 1, $\pm\sqrt{-2}$ or ± 2, and hence show $p(x)$ has no linear factor over F. Show similarly that $p(x)$ is not the product of two quadratics with coefficients in F.]

11. Prove that $x^2 + y^2 - 1$ is irreducible in $\mathbb{Q}[x, y]$.
12. Prove that $x^{n-1} + x^{n-2} + \cdots + x + 1$ is irreducible over $\mathbb{Z}$ if and only if n is a prime.
13. Prove that $x^3 + nx + 2$ is irreducible over $\mathbb{Z}$ for all integers $n \neq 1, -3, -5$.
14. Factor each of the two polynomials: $x^8 - 1$ and $x^6 - 1$ into irreducibles over each of the following rings: (a) $\mathbb{Z}$, (b) $\mathbb{Z}/2\mathbb{Z}$, (c) $\mathbb{Z}/3\mathbb{Z}$.
15. Prove that if F is a field then the polynomial $X^n - x$ which has coefficients in the ring $F[[x]]$ of formal power series (cf. Exercise 3 of Section 7.2) is irreducible over $F[[x]]$. [Recall that $F[[x]]$ is a Euclidean Domain – cf. Exercise 5, Section 7.2 and Example 4, Section 8.1.]
16. Let F be a field and let $f(x)$ be a polynomial of degree n in $F[x]$. The polynomial $g(x) = x^n f(\frac{1}{x})$ is called the *reverse* of $f(x)$.
    (a) Describe the coefficients of g in terms of the coefficients of f.
    (b) Prove that f is irreducible if and only if g is irreducible.
17. Prove the following variant of Eisenstein's Criterion: let P be a prime ideal in the Unique Factorization Domain R and let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial in $R[x]$, $n \geq 1$. Suppose $a_n \notin P$, $a_{n-1}, \ldots, a_0 \in P$ and $a_0 \notin P^2$. Prove that $f(x)$ is irreducible in $F[x]$, where F is the quotient field of R.
18. Show that $6x^5 + 14x^3 - 21x + 35$ and $18x^5 - 30x^2 + 120x + 360$ are irreducible in $\mathbb{Q}[x]$.
19. Let F be a field and let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in F[x]$. The *derivative*, $D_x(f(x))$, of $f(x)$ is defined by

$$D_x(f(x)) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1$$

where, as usual, $na = a + a + \cdots + a$ (n times). Note that $D_x(f(x))$ is again a polynomial with coefficients in F.

The polynomial $f(x)$ is said to have a *multiple root* if there is some field E containing F and some $\alpha \in E$ such that $(x - \alpha)^2$ divides $f(x)$ in $E[x]$. For example, the polynomial $(x-1)^2(x-2) \in \mathbb{Q}[x]$ has $\alpha = 1$ as a multiple root and the polynomial $f(x) = x^4 + 2x^2 + 1 = (x^2+1)^2 \in \mathbb{R}[x]$ has $\alpha = \pm i \in \mathbb{C}$ as multiple roots. We shall prove in Section 13.5 that a nonconstant polynomial $f(x)$ has a multiple root if and only if $f(x)$ is not relatively prime to its derivative (which can be detected by the Euclidean Algorithm in $F[x]$). Use this criterion to determine whether the following polynomials have multiple roots:

    (a) $x^5 + 2x^4 + 4x^3 + 9x^2 + 9x + 3 \in \mathbb{Q}[x]$
    (b) $x^5 - 15x^4 + 85x^3 - 225x^2 + 274x - 120 \in \mathbb{Q}[x]$
    (c) $x^7 - 10x^6 + 27x^5 - 3x^4 - 69x^3 + 189x^2 - 70x + 7 \in \mathbb{Q}[x]$
    (d) Show for any prime p and any $a \in \mathbb{F}_p$ that the polynomial $x^p - a$ has a multiple root.
20. Show that the polynomial $f(x) = x$ in $\mathbb{Z}/6\mathbb{Z}[x]$ factors as $(3x+4)(4x+3)$, hence is not an irreducible polynomial.
    (a) Show that the reduction of $f(x)$ modulo both of the nontrivial ideals (2) and (3) of $\mathbb{Z}/6\mathbb{Z}$ is an irreducible polynomial, showing that the condition that R be an integral domain in Proposition 12 is necessary.
    (b) Show that in any factorization $f(x) = g(x)h(x)$ in $\mathbb{Z}/6\mathbb{Z}[x]$ the reduction of $g(x)$ modulo (2) is either 1 or x and the reduction of $h(x)$ modulo (2) is then either x or 1, and similarly for the reductions modulo (3). Determine all the factorizations of $f(x)$ in $\mathbb{Z}/6\mathbb{Z}[x]$. [Use the Chinese Remainder Theorem.]
    (c) Show that in $\mathbb{Z}/30\mathbb{Z}[x]$ the polynomial $f(x) = x$ has the factorization $f(x) = (10x+21)(15x+16)(6x+25)$. Prove that the product of any of these factors is again of the same degree. Prove that the reduction of $f(x)$ modulo any prime in $\mathbb{Z}/30\mathbb{Z}$ is an irreducible polynomial. Determine all the factorizations of $f(x)$ in $\mathbb{Z}/30\mathbb{Z}[x]$.

[Consider the reductions modulo (2), (3) and (5) and use the Chinese Remainder Theorem.]

(d) Generalize part (c) to $\mathbb{Z}/n\mathbb{Z}[x]$ where n is the product of k distinct primes.

9.5 POLYNOMIAL RINGS OVER FIELDS II

Let F be a field. We prove here some additional results for the one-variable polynomial ring $F[x]$. The first is a restatement of results obtained earlier.

Proposition 15. The maximal ideals in $F[x]$ are the ideals $(f(x))$ generated by irreducible polynomials $f(x)$. In particular, $F[x]/(f(x))$ is a field if and only if $f(x)$ is irreducible.

Proof: This follows from Proposition 8.6 applied to the Principal Ideal Domain $F[x]$.

Proposition 16. Let $g(x)$ be a nonconstant element of $F[x]$ and let

$$g(x) = f_1(x)^{n_1} f_2(x)^{n_2} \dots f_k(x)^{n_k}$$

be its factorization into irreducibles, where the $f_i(x)$ are distinct. Then we have the following isomorphism of rings:

$$F[x]/(g(x)) \cong F[x]/(f_1(x)^{n_1}) \times F[x]/(f_2(x)^{n_2}) \times \cdots \times F[x]/(f_k(x)^{n_k}).$$

Proof: This follows from the Chinese Remainder Theorem (Theorem 7.17), since the ideals $(f_i(x)^{n_i})$ and $(f_j(x)^{n_j})$ are comaximal if $f_i(x)$ and $f_j(x)$ are distinct (they are relatively prime in the Euclidean Domain $F[x]$, hence the ideal generated by them is $F[x]$).

The next result concerns the number of roots of a polynomial over a field F. By Proposition 9, a root α corresponds to a linear factor $(x - \alpha)$ of $f(x)$. If $f(x)$ is divisible by $(x - \alpha)^m$ but not by $(x - \alpha)^{m+1}$, then α is said to be a root of *multiplicity* m.

Proposition 17. If the polynomial $f(x)$ has roots $\alpha_1, \alpha_2, \dots, \alpha_k$ in F (not necessarily distinct), then $f(x)$ has $(x - \alpha_1) \cdots (x - \alpha_k)$ as a factor. In particular, a polynomial of degree n in one variable over a field F has at most n roots in F, even counted with multiplicity.

Proof: The first statement follows easily by induction from Proposition 9. Since linear factors are irreducible, the second statement follows since $F[x]$ is a Unique Factorization Domain.

This last result has the following interesting consequence:

Proposition 18. A finite subgroup of the multiplicative group of a field is cyclic. In particular, if F is a finite field, then the multiplicative group $F^\times$ of nonzero elements of F is a cyclic group.

Proof: We give a proof of this result using the Fundamental Theorem of Finitely Generated Abelian Groups (Theorem 3 in Section 5.2). A more number-theoretic proof is outlined in the exercises, or Proposition 5 in Section 6.1 may be used in place of the Fundamental Theorem. By the Fundamental Theorem, the finite subgroup can be written as the direct product of cyclic groups

$$\mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z} \times \cdots \times \mathbb{Z}/n_k\mathbb{Z}$$

where $n_k \mid n_{k-1} \mid \cdots \mid n_2 \mid n_1$. In general, if G is a cyclic group and $d \mid |G|$ then G contains precisely d elements of order dividing d. Since n_k divides the order of each of the cyclic groups in the direct product, it follows that each direct factor contains n_k elements of order dividing n_k. If k were greater than 1, there would therefore be a total of more than n_k such elements. But then there would be more than n_k roots of the polynomial $x^{n_k} - 1$ in the field F, contradicting Proposition 17. Hence $k = 1$ and the group is cyclic.

Corollary 19. Let p be a prime. The multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ of nonzero residue classes mod p is cyclic.

Proof: This is the multiplicative group of the finite field $\mathbb{Z}/p\mathbb{Z}$.

Corollary 20. Let $n \geq 2$ be an integer with factorization $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$ in $\mathbb{Z}$, where $p_1, \ldots, p_r$ are distinct primes. We have the following isomorphisms of (multiplicative) groups:

(1) $(\mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/p_1^{\alpha_1}\mathbb{Z})^\times \times (\mathbb{Z}/p_2^{\alpha_2}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_r^{\alpha_r}\mathbb{Z})^\times$

(2) $(\mathbb{Z}/2^\alpha\mathbb{Z})^\times$ is the direct product of a cyclic group of order 2 and a cyclic group of order $2^{\alpha-2}$, for all $\alpha \geq 2$

(3) $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ is a cyclic group of order $p^{\alpha-1}(p-1)$, for all odd primes p.

Remark: These isomorphisms describe the group-theoretic structure of the automorphism group of the cyclic group, Z_n, of order n since $\mathrm{Aut}(Z_n) \cong (\mathbb{Z}/n\mathbb{Z})^\times$ (cf. Proposition 16 in Section 4.4). In particular, for p a prime the automorphism group of the cyclic group of order p is cyclic of order $p - 1$.

Proof: This is mainly a matter of collecting previous results. The isomorphism in (1) follows from the Chinese Remainder Theorem (see Corollary 18, Section 7.6). The isomorphism in (2) follows directly from Exercises 22 and 23 of Section 2.3.

For p an odd prime, $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ is an abelian group of order $p^{\alpha-1}(p-1)$. By Exercise 21 of Section 2.3 the Sylow p-subgroup of this group is cyclic. The map

$$\mathbb{Z}/p^\alpha\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z} \qquad \text{defined by} \qquad a + (p^\alpha) \mapsto a + (p)$$

is a ring homomorphism (reduction mod p) which gives a surjective group homomorphism from $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ onto $(\mathbb{Z}/p\mathbb{Z})^\times$. The latter group is cyclic of order $p - 1$ (Corollary 19). The kernel of this map is of order $p^{\alpha-1}$, hence for all primes $q \neq p$, the Sylow q-subgroup of $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ maps isomorphically into the cyclic group $(\mathbb{Z}/p\mathbb{Z})^\times$. All Sylow subgroups of $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ are therefore cyclic, so (3) holds, completing the proof.

One interesting arithmetic consequence of Corollary 19 is the following:

Corollary 21. The prime number p divides an integer of the form $n^2 + 1$ if and only if p is either 2 or an odd prime congruent to 1 modulo 4.

Proof: The statement for $p = 2$ is trivial since $2 \mid 1^2 + 1$. If p is an odd prime, note that $p \mid n^2 + 1$ is equivalent to $n^2 = -1$ in $\mathbb{Z}/p\mathbb{Z}$. This in turn is equivalent to saying the residue class of n is of order 4 in the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$. Thus p divides an integer of the form $n^2 + 1$ if and only if $(\mathbb{Z}/p\mathbb{Z})^\times$ contains an element of order 4. By Corollary 19, this group is *cyclic*, hence it contains an element of order 4 if and only if 4 divides its order, $p - 1$, i.e. $p \equiv 1 \pmod 4$.

Using the fact that the ring $\mathbb{Z}[i]$ is Euclidean, hence a Principal Ideal Domain, the result above can be used to show that the conditions in Corollary 21 are also equivalent to p being *the sum of two (integer) squares*: $p = a^2 + b^2$. This is a famous theorem due to Fermat (see the exercises).

EXERCISES

1. Let F be a field and let $f(x)$ be a nonconstant polynomial in $F[x]$. Describe the nilradical of $F[x]/(f(x))$ in terms of the factorization of $f(x)$ (cf. Exercise 29, Section 7.3).
2. For each of the fields constructed in Exercise 6 of Section 4 exhibit a generator for the (cyclic) multiplicative group of nonzero elements.
3. This problem completes the proof of *Fermat's Theorem*: a prime p is the sum of two integer squares if and only if it equals 2 or is congruent to 1 mod 4. The case of $p = 2 = 1^2 + 1^2$ is clear, so let p denote an odd prime. Recall that the norm of $a + bi$ is $N(a + bi) = a^2 + b^2$.
   (a) Show that if $p = a^2 + b^2$ $(a, b \in \mathbb{Z})$ then $p \equiv 1$ mod 4.
   (b) Suppose now that p is a prime in $\mathbb{Z}$ with $p \equiv 1$ mod 4. Then Corollary 21 shows that p divides an integer of the form $n^2 + 1$. Factor $n^2 + 1$ in the Euclidean domain $\mathbb{Z}[i]$ (cf. Exercise 7, Section 8.1) and show that the assumption that p is a prime (= irreducible) in $\mathbb{Z}[i]$ leads to a contradiction.
   (c) From (b) it follows that p is divisible by some $a + bi$ which is not a unit. By computing norms, argue that then $p = (a + bi)(a - bi)$, showing that p is the sum of two squares.
   (d) Show that a number which is a prime in $\mathbb{Z}$ factors in $\mathbb{Z}[i]$ as follows: 2 is a unit times the square of the prime $1 + i$; if $p \equiv 3$ mod 4 then p remains prime in $\mathbb{Z}[i]$; if $p \equiv 1$ mod 4 then p is the product of two distinct primes in $\mathbb{Z}[i]$.
4. This problem extends the results of the previous exercise to prove more generally that a positive integer n is the sum of two squares if and only if the prime divisors of n that are congruent to 3 mod 4 occur to even exponents. We first determine all the irreducibles (= primes) in $\mathbb{Z}[i]$.
   (a) Let π be an irreducible in $\mathbb{Z}[i]$ and let (π) be the ideal generated by π in $\mathbb{Z}[i]$. Prove that $(\pi) \cap \mathbb{Z}$ is a prime ideal in $\mathbb{Z}$, say $p\mathbb{Z}$. Conclude that π divides some integer prime p.
   (b) Prove that if $p = 2$ then $\pi = 1 + i$ (up to associates), with $N(1 + i) = 2$ (here N is the usual norm on $\mathbb{Z}[i]$).

   Prove that if $p \equiv 3$ mod 4 then $\pi = p$ (up to associates) is irreducible in $\mathbb{Z}[i]$, with $N(\pi) = p^2$ in this case.

   Prove that if $p \equiv 1$ mod 4 then $\pi = a \pm bi$ (up to associates), where $p = N(\pi) = a^2 + b^2$ as in the previous exercise.

(c) Prove that the integer n is the sum of two integer squares if and only if every prime $q \equiv 3 \bmod 4$ occurring in the prime power factorization of n occurs to an even exponent. [The integer n is the sum of two squares if and only if n is the norm of some element in $\mathbb{Z}[i]$. Write n as the unique product of primes in the Unique Factorization Domain $\mathbb{Z}[i]$ and use (b).]

(d) Use (c) to show $493 = 17 \cdot 29$ is the sum of two squares. [Write $17 = (4+i)(4-i)$ and $29 = (5+2i)(5-2i)$ to see that $493 = (4+i)(4-i)(5+2i)(5-2i)$ is the prime factorization of 493 in $\mathbb{Z}[i]$. Since $(4+i)(5+2i) = 18+13i$, this gives $493 = (18+13i)(18-13i) = 18^2 + 13^2$.] As this example illustrates, the entire question of writing integers as the sum of two squares is nothing more than a factorization question in the Unique Factorization Domain $\mathbb{Z}[i]$.

5. Prove that if an integer is the sum of two rational squares, then it is the sum of two integer squares (for example, $13 = (\frac{1}{5})^2 + (\frac{18}{5})^2 = 2^2 + 3^2$).

6. Let p be an odd prime in $\mathbb{Z}$ and let n be a positive integer. Prove that $x^n - p$ is irreducible over $\mathbb{Z}[i]$. [Use Exercise 3 and Eisenstein's Criterion.]

7. Prove that $x^3 + 12x^2 + 18x + 6$ is irreducible over $\mathbb{Z}[i]$. [Use Exercise 3 and Eisenstein's Criterion.]

8. Show that $\mathbb{Z}[i]/(3)$ is a field with 9 elements.

9. Let φ denote Euler's φ–function. Prove the identity $\sum_{d|n} \varphi(d) = n$, where the sum is extended over all the divisors d of n. [First observe that the identity is valid when $n = p^m$ is the power of a prime p since the sum telescopes. Write $n = p^m n'$ where p does not divide n'. Prove that $\sum_{d|n} \varphi(d) = \sum_{d''|p^m} \varphi(d'') \sum_{d'|n'} \varphi(d')$ by multiplying out the right hand side and using the multiplicativity $\varphi(ab) = \varphi(a)\varphi(b)$ when a and b are relatively prime. Use induction to complete the proof. This problem may be done alternatively by letting Z be the cyclic group of order n and showing that since Z contains a unique subgroup of order d for each d dividing n, the number of elements of Z of order d is $\varphi(d)$. Then $|Z|$ is the sum of $\varphi(d)$ as d runs over all divisors of n.]

10. Let G be a finite subgroup of order n of the multiplicative group $F^\times$ of nonzero elements of the field F. Let φ denote Euler's φ–function and let $\psi(d)$ denote the number of elements of G of order d. Prove that $\psi(d) = \varphi(d)$ for every divisor d of n. In particular conclude that $\psi(n) \geq 1$, so that G is a cyclic group. [Observe that for any integer $N \geq 1$ the polynomial $x^N - 1$ has at most N roots in F. Conclude that for any integer N we have $\sum_{d|N} \psi(d) \leq N$. Since $\sum_{d|N} \varphi(d) = N$ by the previous exercise, show by induction that $\psi(d) \leq \varphi(d)$ for every divisor d of n. Since $\sum_{d|n} \psi(d) = n = \sum_{d|n} \varphi(d)$ show that this implies $\psi(d) = \varphi(d)$ for every divisor d of n.]

11. Prove that the additive and multiplicative groups of a field are never isomorphic.

For the remaining exercises let F be a field, let F^n be the set of all n-tuples of elements of F (called *affine n-space* over F) and let R be the polynomial ring $F[x_1, x_2, \ldots, x_n]$. The elements of R form a ring of F-valued functions on F^n, where the value of the polynomial $p(x_1, \ldots, x_n)$ on the n-tuple $(a_1, \ldots, a_n)$ is obtained by substituting a_i for x_i for all i (so if $F = \mathbb{R}$ this is the usual way of evaluating a function of n variables at a point in Euclidean n-space, $\mathbb{R}^n$). These exercises explore correspondences between subsets of F^n and ideals in R analogous to those discussed in Exercises 34 to 37 of Section 7.4. Once again we shall see that, in certain cases, the points (and a topology) on the space F^n can be described completely in terms of the ideal theory of the ring R, so that geometric properties and problems in F^n can be

translated into ring-theoretic terms in R. In this way, problems involving polynomials such as questions about the zeros of polynomials (e.g. the zeros of the celebrated Fermat polynomial $x^p + y^p - z^p$) may be studied by ring theory. As mentioned in the exercises in Section 7.4, this process is the basis for a vast area of mathematics called *algebraic geometry*. One advantage of the current "algebraic" (polynomial) setting over its "continuous" analog discussed in Section 7.4 is that the ring $F[x_1, \ldots, x_n]$ is *Noetherian*, i.e. every ideal is finitely generated (cf. Exercise 8 of Section 1).

Proceeding as in Exercises 34 to 37 in Section 7.4, define maps $I : \{\text{ideals of } R\} \to \{\text{subsets of } F^n\}$ and $V : \{\text{subsets of } F^n\} \to \{\text{ideals of } R\}$ as follows. For each subset X of F^n let $I(X)$ be the subset of R consisting of all functions that vanish on X:

$$I(X) = \{f \in R \mid f(a) = 0 \text{ for all } a \in X\},$$

with the convention that $I(\emptyset) = R$. For each subset J of R let $V(J)$ be the subset of F^n consisting of the points where all functions in J simultaneously vanish — called the *locus* of J:

$$V(J) = \{a = (a_1, \ldots, a_n) \in F^n \mid f(a) = 0 \text{ for all } f \in J\},$$

with the convention that $V(\emptyset) = F^n$. Note that if $n = 1$, the locus of a polynomial just consists of the *roots* of the polynomial that lie in F.

12. Establish the following properties of I and V.

(a) Prove that $I(X)$ is an ideal of R for all subsets X of F^n (hence the use of the letter I).
Prove that if J is any subset of R and (J) is the ideal generated by J in R, then $V(J) = V((J))$ (so it suffices to restrict attention to ideals J in R when studying V).

(b) Prove that if $X \subseteq Y \subseteq F^n$ then $I(Y) \subseteq I(X)$.
Prove that if $I \subseteq J$ are ideals of R then $V(J) \subseteq V(I)$.

(c) Prove that $I(X \cup Y) = I(X) \cap I(Y)$ for all $X, Y \subseteq F^n$.
Prove that if I and J are ideals of R then $V(I + J) = V(I) \cap V(J)$.
Prove that if I and J are ideals of R then $V(IJ) = V(I) \cup V(J)$ (where IJ is the ideal which is the product of I and J — cf. Exercise 34, Section 7.3).

(d) Prove that $X \subseteq V(I(X))$ for all subsets X of F^n.
Prove that $J \subseteq I(V(J))$ for all ideals J of R.

If c is a point in affine n-space F^n, the ideal $I(\{c\}) = \{f \in R \mid f(c) = 0\}$ (denoted M_c in Section 7.4) is a maximal ideal of R since it is the kernel of the surjective homomorphism from R to the field F given by evaluation at c. In general, not every maximal ideal of R is of the form $I(\{c\})$ for some $c \in F^n$. For example, let $n = 1$ and $F = \mathbb{R}$. Then the ideal $(x^2 + 1)$ is maximal in $\mathbb{R}[x]$ (the quotient ring is the field $\mathbb{C}$) but there is no c such that $(x^2 + 1) = I(\{c\})$ (because $x^2 + 1$ has no zeros in affine 1-space, $\mathbb{R}$). It turns out that this example describes the generic reason why the V and I maps are not bijections between the points of F^n and the maximal ideals of R, in the following sense. A field F is said to be *algebraically closed* if every nonconstant polynomial $f(x) \in F[x]$ has a root in F, i.e. there is some $c \in F$ such that $f(c) = 0$. (We shall study algebraically closed fields in Section 13.4; we shall also prove, in Section 14.6, that the complex numbers, $\mathbb{C}$, is an algebraically closed field.)

Theorem. (Hilbert's Nullstellensatz, weak form) If F is an algebraically closed field then for every maximal ideal M of $R = F[x_1, \ldots, x_n]$ there is a point $c \in F^n$ such that $M = I(\{c\}) = \{f \in R \mid f(c) = 0\}$.

Proofs of Hilbert's Nullstellensatz may be found in most books on algebraic curves, algebraic geometry or commutative ring theory.

13. Let F be an algebraically closed field. Assuming Hilbert's Nullstellensatz, prove that the maps I and V are bijections between the set of points of F^n and the set of maximal ideals of R and that V and I are inverses of each other on these sets.

Assume now that F is an algebraically closed field. As in the case of the ring of continuous functions on the closed interval [0,1], there is a topology on F^n such that the *closed sets* are in bijective correspondence with certain *ideals* of R. More precisely, a subset X of F^n is called an *affine algebraic set* if $X = V(J)$ for some ideal J of R. It follows from the exercises above that the affine algebraic sets satisfy the axioms for a collection of closed sets in a topological space (i.e. the empty set and the whole space are closed sets, a finite union of closed sets is closed and an arbitrary intersection of closed sets is closed). The topology on F^n obtained by taking the affine algebraic sets as the closed sets is called the *Zariski topology* on F^n. When $n = 1$, every polynomial has only a finite number of zeros in F and so the affine algebraic subsets of F are precisely the finite subsets of F. It is easy to see that the Zariski topology on F^n is the coarsest topology that makes all the elements of R continuous (considered as maps from F^n to F, where the closed sets in F are the finite sets). Recall (Exercise 30, Section 7.4) that the radical of an ideal J of R is defined as

$$\mathcal{R}(J) = \{f \in R \mid f^k \in J \text{ for some } k \in \mathbb{Z}^+\}$$

and an ideal J is called a radical ideal if $J = \mathcal{R}(J)$. The next theorem shows that the radical ideals play the role analogous to the ideals J in Exercise 36 of Section 7.4 that satisfied $J = \mathcal{J}(J)$.

Theorem. (Hilbert's Nullstellensatz, strong form) If F is an algebraically closed field then $I(V(J)) = \mathcal{R}(J)$ for all ideals J of R.

This theorem, which is usually derived from its weak form stated above, can also be readily found in the literature.

14. Let F be an algebraically closed field. Assuming Hilbert's Nullstellensatz (strong form), prove that the maps I and V are bijections between the collection of affine algebraic subsets of F^n and the collection of radical ideals of R, and that V and I are inverses of each other on these sets.

Finally, we give two exercises which illustrate the interplay between the topological and ring-theoretic objects introduced.

15. For each $f \in R$ let $V_f = V((f))$ be the set of zeros of f (which is an affine algebraic subset of F^n). Show that if X is an affine algebraic set then there are a finite number of

polynomials $f_1, \ldots, f_m \in R$ such that $X = V_{f_1} \cap \cdots \cap V_{f_m}$. [Use the fact that R is Noetherian.]

16. Assume F is an algebraically closed field. An affine algebraic set X is said to be *irreducible* if one cannot write $X = X_1 \cup X_2$ for some affine algebraic sets X_1, X_2 which are both proper subsets of X. Prove that X is irreducible if and only if $I(X)$ is a prime ideal of R.

Part III

MODULES AND VECTOR SPACES

In Part III we study the mathematical objects called modules. The use of modules was pioneered by one of the most prominent mathematicians of the first part of this century, Emmy Noether, who led the way in demonstrating the power and elegance of this structure. We shall see that vector spaces are just special types of modules which arise when the underlying ring is a field. If R is a ring, the definition of an R-module M is closely analogous to the definition of a group action where R plays the role of the group and M the role of the set. The additional axioms for a module require that M itself has more structure (namely that M be an abelian group). Modules are the "representation objects" for rings, i.e. they are, by definition, algebraic objects on which rings act. As the theory develops it will become apparent how the structure of the ring R (in particular, the structure and wealth of its ideals) is reflected by the structure of its modules and vice versa in the same way that the structure of the collection of normal subgroups of a group was reflected by its permutation representations.

10

Introduction to Module Theory

10.1 BASIC DEFINITIONS AND EXAMPLES

We start with the definition of a module.

Definition. Let R be a ring (not necessarily commutative nor with 1). A *left R-module* or a *left module over R* is a set M together with

(1) a binary operation + on M under which M is an abelian group, and

(2) an action of R on M (that is, a map $R \times M \to M$) denoted by rm, for all $r \in R$ and for all $m \in M$ which satisfies

(a) $(r + s)m = rm + sm$, for all $r, s \in R$, $m \in M$,

(b) $(rs)m = r(sm)$, for all $r, s \in R$, $m \in M$, and

(c) $r(m + n) = rm + rn$, for all $r \in R$, $m, n \in M$.

If the ring R has a 1 we impose the additional axiom:

(d) $1m = m$, for all $m \in M$.

The descriptor "left" in the above definition indicates that the ring elements appear on the left; "right" R-modules can be defined analogously. If the ring R is *commutative* and M is a left R-module we can make M into a right R-module by defining $mr = rm$ for $m \in M$ and $r \in R$. If R is not commutative, axiom 2(b) in general will not hold with this definition (so not every left R-module is also a right R-module). Unless explicitly mentioned otherwise the term "module" will always mean "left module." Modules satisfying axiom 2(d) are called *unital* modules and in this book all our modules will be unital (this is to avoid "pathologies" such as having $rm = 0$ for all $r \in R$ and $m \in M$).

When R is a field F the axioms for an R-module are precisely the same as those for a vector space over F, so that

modules over a field F and vector spaces over F are the same.

Before giving other examples of R-modules we record the obvious definition of submodules.

Definition. Let R be a ring and let M be an R-module. An *R-submodule* of M is a subgroup N of M which is closed under the action of ring elements, i.e. $rn \in N$, for all $r \in R$, $n \in N$.

Submodules of M are therefore just subsets of M which are themselves modules under the restricted operations. In particular, if $R = F$ is a field, submodules are the same as subspaces. Every R-module M has the two submodules M and 0 (the latter is called the *trivial submodule*).

Examples

(1) Let R be any ring. Then $M = R$ is a left R-module, where the action of a ring element on a module element is just the usual multiplication in the ring R (similarly, R is a right module over itself). In particular, every field can be considered as a (1-dimensional) vector space over itself. When R is considered as a left module over itself in this fashion, the submodules of R are precisely the left ideals of R (and if R is considered as a right R-module over itself, its submodules are the right ideals). Thus if R is not commutative it has a left and right module structure over itself and these structures may be different (e.g. the submodules may be different) — an exercise at the end of this section gives a specific example of this.

(2) Let $R = F$ be a field. As noted above, every vector space over F is an F-module and vice versa. Let $n \in \mathbb{Z}^+$ and let

$$F^n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in F, \text{ for all } i\}$$

(called *affine n-space over F*). Make F^n into a vector space by defining addition and scalar multiplication componentwise:

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)$$
$$\alpha(a_1, \ldots, a_n) = (\alpha a_1, \ldots, \alpha a_n), \qquad \alpha \in F.$$

As in the case of Euclidean n-space (i.e. when $F = \mathbb{R}$), affine n-space is a vector space of dimension n over F (we shall discuss the notion of dimension more thoroughly in the next chapter).

(3) Let R be a ring with 1 and let $n \in \mathbb{Z}^+$. Following Example 2 define

$$R^n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in R, \text{ for all } i\}.$$

Make R^n into an R-module by componentwise addition and multiplication by elements of R in the same manner as when R was a field. The module R^n is called *the free module of rank n over R*. (We shall see shortly that free modules have the same "universal property" in the context of R-modules that free groups were seen to have in Section 6.3. We shall also soon discuss direct products of R-modules.) An obvious submodule of R^n is given by the ith component, namely the set of n-tuples with arbitrary ring elements in the ith component and zeros in the jth component for all $j \neq i$.

(4) The same abelian group may have the structure of an R-module for a number of different rings R and each of these module structures may carry useful information. Specifically, if M is an R-module and S is a subring of R, then M is automatically an S-module as well. For instance the field $\mathbb{R}$ is an $\mathbb{R}$-module, a $\mathbb{Q}$-module and a $\mathbb{Z}$-module.

(5) If M is an R-module and for some (2-sided) ideal I of R, $am = 0$, for all $a \in I$ and all $m \in M$, we say M is *annihilated by I*. In this situation we can make M into an (R/I)-module by defining an action of the quotient ring R/I on M as follows: for each $m \in M$ and coset $r + I$ in R/I let

$$(r + I)m = rm.$$

Since $am = 0$ for all $a \in I$ and all $m \in M$ this is well defined and one easily checks that it makes M into an (R/I)-module. In particular, when I is a maximal ideal in the commutative ring R and $IM = 0$, then M is a vector space over the field R/I (cf. the following example).

The next example is of sufficient importance as to be singled out. It will form the basis for our proof of the Fundamental Theorem of Finitely Generated Abelian Groups in Chapter 12.

Example: ($\mathbb{Z}$-modules)

Let R $= \mathbb{Z}$, let A be *any* abelian group (finite or infinite) and write the operation of A as $+$. Make A into a $\mathbb{Z}$-module as follows: for any $n \in \mathbb{Z}$ and $a \in A$ define

$$na = \begin{cases} a + a + \cdots + a \ \ (n \text{ times}) & \text{if } n > 0 \\ 0 & \text{if } n = 0 \\ -a - a - \cdots - a \ \ (-n \text{ times}) & \text{if } n < 0 \end{cases}$$

(here 0 is the identity of the additive group A). This definition of an action of the integers on A makes A into a $\mathbb{Z}$-module. Thus every abelian group is a $\mathbb{Z}$-module. Conversely, if M is any $\mathbb{Z}$-module, a fortiori M is an abelian group, so

$\mathbb{Z}$-modules are the same as abelian groups.

Furthermore, it is immediate from the definition that

$\mathbb{Z}$-submodules are the same as subgroups.

Note that for the cyclic group $\langle a \rangle$ written multiplicatively the additive notation na becomes a^n, that is, we have all along been using the fact that $\langle a \rangle$ is a right $\mathbb{Z}$-module (checking that this "exponential" notation satisfies the usual laws of exponents is equivalent to checking the $\mathbb{Z}$-module axioms — this was given as an exercise at the end of Section 1.1). Note that since $\mathbb{Z}$ is commutative these definitions of left and right actions by ring elements give the same module structure.

If A is an abelian group containing an element x of finite order n then $nx = 0$. Thus, in contrast to vector spaces, a $\mathbb{Z}$-module may have nonzero elements x such that $nx = 0$ for some nonzero ring element n. In particular, if A has order m, then by Lagrange's Theorem (Corollary 9, Section 3.2) $mx = 0$, for *all* $x \in A$. Note that then A is a module over $\mathbb{Z}/m\mathbb{Z}$.

In particular, if p is a prime and A is an abelian group (written additively) such that $px = 0$, for all $x \in A$, then (as noted in Example 5) A is a $\mathbb{Z}/p\mathbb{Z}$-module, i.e. can be considered as a vector space over the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. For instance, the Klein 4-group is a (2-dimensional) vector space over $\mathbb{F}_2$. These groups are the *elementary abelian p-groups* discussed in Section 4.4 (see, in particular, Proposition 17(3)).

The next example is also of fundamental importance and will form the basis for our study of canonical forms of matrices in Sections 12.2 and 12.3.

Example: (F[x]-modules)

Let F be a field, let x be an indeterminate and let R be the polynomial ring $F[x]$. Let V be a vector space over F and let T be a linear transformation from V to V (we shall review the theory of linear transformations in the next chapter — for the purposes of this example one only needs to know the definition of a linear transformation). We have

already seen that V is an F-module; the linear map T will enable us to make V into an $F[x]$-module.

First, for the nonnegative integer n, define

$$\begin{aligned} T^0 &= I, \\ &\vdots \\ T^n &= T \circ T \circ \cdots \circ T \qquad (n \text{ times}) \end{aligned}$$

where I is the identity map from V to V and $\circ$ denotes function composition (which makes sense because the domain and codomain of T are the same). Also, for any two linear transformations A, B from V to V and elements $\alpha, \beta \in F$, let $\alpha A + \beta B$ be defined by

$$(\alpha A + \beta B)(v) = \alpha(A(v)) + \beta(B(v))$$

(i.e. addition and scalar multiplication of linear transformations are defined pointwise). Then $\alpha A + \beta B$ is easily seen to be a linear transformation from V to V, so that linear combinations of linear transformations are again linear transformations.

We now define the action of any polynomial in x on V. Let $p(x)$ be the polynomial

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_0, \ldots, a_n \in F$. For each $v \in V$ define an action of the ring element $p(x)$ on the module element v by

$$\begin{aligned} p(x)v &= (a_n T^n + a_{n-1}T^{n-1} + \cdots + a_1 T + a_0)(v) \\ &= a_n T^n(v) + a_{n-1}T^{n-1}(v) + \cdots + a_1 T(v) + a_0 v \end{aligned}$$

(i.e. $p(x)$ acts by substituting the linear transformation T for x in $p(x)$ and applying the resulting linear transformation to v). Put another way, x acts on V as the linear transformation T and we extend this to an action of all of $F[x]$ on V in a natural way. It is easy to check that this definition of an action of $F[x]$ on V satisfies all the module axioms and makes V into an $F[x]$-module.

The field F is naturally a subring of $F[x]$ (the constant polynomials) and the action of these field elements is by definition the same as their action when viewed as constant polynomials. In other words, the definition of the $F[x]$ action on V is consistent with the given action of the field F on the vector space V, i.e. the definition *extends* the action of F to an action of the larger ring $F[x]$.

The way $F[x]$ acts on V depends on the choice of T so that there are in general many different $F[x]$-module structures on the same vector space V. For instance, if $T = 0$, and $p(x), v$ are as above, then $p(x)v = a_0 v$, that is, the polynomial $p(x)$ acts on v simply by multiplying by the constant term of $p(x)$, so that the $F[x]$-module structure is just the F-module structure. If, on the other hand, T is the identity transformation (so $T^n(v) = v$, for all n and v), then $p(x)v = a_n v + a_{n-1}v + \cdots + a_0 v = (a_n + \cdots + a_0)v$, so that now $p(x)$ multiplies v by the sum of the coefficients of $p(x)$.

To give another specific example, let V be affine n-space F^n and let T be the "shift operator"

$$T(x_1, x_2, \ldots, x_n) = (x_2, x_3, \ldots, x_n, 0).$$

Let e_i be the usual i^{th} basis vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ where the 1 is in position i. Then

$$T^k(e_i) = \begin{cases} e_{i-k} & \text{if } i > k \\ 0 & \text{if } i \le k \end{cases}$$

so for example, if $m < n$,

$$(a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0)e_n = (0, \ldots, 0, a_m, a_{m-1}, \ldots, a_0).$$

From this we can determine the action of any polynomial on any vector.

The construction of an $F[x]$-module from a vector space V over F and a linear transformation T from V to V in fact describes *all* $F[x]$-modules; namely, an $F[x]$-module is a vector space together with a linear transfomation which specifies the action of x. This is because if V is any $F[x]$-module, then V is an F-module and the action of the ring element x on V is a linear transformation from V to V. The axioms for a module ensure that the actions of F and x on V uniquely determine the action of any element of $F[x]$ on V. Thus there is a bijection between the collection of $F[x]$-modules and the collection of pairs V, T

$$\left\{ V \text{ an } F[x]\text{-module} \right\} \longleftrightarrow \left\{ \begin{array}{c} V \text{ a vector space over } F \\ \text{and} \\ T : V \to V \text{ a linear transformation} \end{array} \right\}$$

given by

the element x acts on V as the linear transformation T.

Now we consider $F[x]$-submodules of V where, as above, V is any $F[x]$-module and T is the linear transformation from V to V given by the action of x. An $F[x]$-submodule W of V must first be an F-submodule, i.e. W must be a vector subspace of V. Secondly, W must be sent to itself under the action of the ring element x, i.e. we must have $T(w) \in W$, for all $w \in W$. Any vector subspace U of V such that $T(U) \subseteq U$ is called *T-stable* or *T-invariant*. If U is any T-stable subspace of V it follows that $T^n(U) \subseteq U$, for all $n \in \mathbb{Z}^+$ (for example, $T(U) \subseteq U$ implies $T^2(U) = T(T(U)) \subseteq T(U) \subseteq U$). Moreover any linear combination of powers of T then sends U into U so that U is also stable by the action of *any* polynomial in T. Thus U is an $F[x]$-submodule of V. This shows that

the $F[x]$-submodules of V are precisely the T-stable subspaces of V.

In terms of the bijection above:

$$\left\{ W \text{ an } F[x]\text{-submodule} \right\} \longleftrightarrow \left\{ \begin{array}{c} W \text{ a subspace of } V \\ \text{and} \\ W \text{ is } T\text{-stable} \end{array} \right\}$$

which gives a complete dictionary between $F[x]$-modules V and vector spaces V together with a given linear transformation T from V to V.

For instance, if T is the shift operator defined on affine n-space above and k is any integer in the range $0 \leq k \leq n$, then the subspace

$$U_k = \{ (x_1, x_2, \ldots, x_k, 0, \ldots, 0) \mid x_i \in F \}$$

is clearly T-stable so is an $F[x]$-submodule of V.

We emphasize that an abelian group M may have many different R-module structures, even if the ring R does not vary (in the same way that a given group G may act in many ways as a permutation group on some fixed set Ω). We shall see that the structure of an R-module is reflected by the ideal structure of R. When R is a field (the subject of the next chapter) all R-modules will be seen to be products of copies of R (as in Example 3 above).

We shall see in Chapter 12 that the relatively simple ideal structure of the ring $F[x]$ (recall that $F[x]$ is a Principal Ideal Domain) forces the $F[x]$-module structure of V to be correspondingly uncomplicated, and this in turn provides a great deal of information about the linear transformation T (in particular, gives some nice matrix representations for T: its *rational canonical form* and its *Jordan canonical form*). Moreover, the same arguments which *classify* finitely generated $F[x]$-modules apply to any Principal Ideal Domain R, and when these are invoked for $R = \mathbb{Z}$, we obtain the Fundamental Theorem of Finitely Generated Abelian Groups. These results generalize the theorem that every finite dimensional vector space has a basis.

In Part V of the book we shall study modules over certain noncommutative rings (group rings) and see that this theory in some sense generalizes both the study of $F[x]$-modules in Chapter 12 and the notion of a permutation representation of a finite group.

We close this section with a submodule criterion analogous to that for subgroups of a group in Section 2.1.

Proposition 1. (The Submodule Criterion) Let R be a ring and let M be an R-module. A subset N of M is a submodule of M if and only if

(1) $N \neq \emptyset$, and

(2) $x + \alpha y \in N$ for all $\alpha \in R$ and for all $x, y \in N$.

Proof: If N is a submodule, then $0 \in N$ so $N \neq \emptyset$. Also N is closed under addition and is sent to itself under the action of elements of R. Conversely, suppose (1) and (2) hold. Let $\alpha = -1$ and apply the subgroup criterion (in additive form) to see that N is a subgroup of M. In particular, $0 \in N$. Now let $x = 0$ and apply hypothesis (2) to see that N is sent to itself under the action of R. This establishes the proposition.

EXERCISES

In these exercises R is a ring with 1 and M is a left R-module.

1. Prove that $0m = 0$ and $(-1)m = -m$ for all $m \in M$.
2. Prove that $R^\times$ and M satisfy the two axioms in Section 1.7 for a *group action* of the multiplicative group $R^\times$ on the set M.
3. Assume that $rm = 0$ for some $r \in R$ and some $m \in M$ with $m \neq 0$. Prove that r does not have a left inverse (i.e. there is no $s \in R$ such that $sr = 1$).
4. Let M be the module R^n described in Example 3 and let $I_1, I_2, \ldots, I_n$ be left ideals of R. Prove that the following are submodules of M:
   (a) $\{(x_1, x_2, \ldots, x_n) \mid x_i \in I_i\}$
   (b) $\{(x_1, x_2, \ldots, x_n) \mid x_i \in R \text{ and } x_1 + x_2 + \cdots + x_n = 0\}$.
5. For any left ideal I of R define

$$IM = \{\sum_{\text{finite}} a_i m_i \mid a_i \in I,\ m_i \in M\}$$

to be the collection of all finite sums of elements of the form am where $a \in I$ and $m \in M$. Prove that IM is a submodule of M.

6. Show that the intersection of any nonempty collection of submodules of an R-module is a submodule.

7. Let $N_1 \subseteq N_2 \subseteq \cdots$ be an ascending chain of submodules of M. Prove that $\cup_{i=1}^{\infty} N_i$ is a submodule of M.

8. An element m of the R-module M is called a *torsion element* if $rm = 0$ for some nonzero element $r \in R$. The set of torsion elements is denoted

$$\mathrm{Tor}(M) = \{m \in M \mid rm = 0 \text{ for some nonzero } r \in R\}.$$

   (a) Prove that if R is an integral domain then $\mathrm{Tor}(M)$ is a submodule of M (called the *torsion* submodule of M).
   (b) Give an example of a ring R and an R-module M such that $\mathrm{Tor}(M)$ is not a submodule. [Consider the torsion elements in the R-module R.]
   (c) Show that if R has zero divisors then every nonzero R-module has torsion elements.

9. If N is a submodule of M, the *annihilator of N in R* is defined to be $\{r \in R \mid rn = 0 \text{ for all } n \in N\}$. Prove that the annihilator of N in R is a 2-sided ideal of R.

10. If I is a right ideal of R, the *annihilator of I in M* is defined to be $\{m \in M \mid am = 0 \text{ for all } a \in I\}$. Prove that the annihilator of I in M is a submodule of M.

11. Let M be the abelian group (i.e. $\mathbb{Z}$-module) $\mathbb{Z}/24\mathbb{Z} \times \mathbb{Z}/15\mathbb{Z} \times \mathbb{Z}/50\mathbb{Z}$.
   (a) Find the annihilator of M in $\mathbb{Z}$ (i.e. a generator for this principal ideal).
   (b) Let $I = 2\mathbb{Z}$. Describe the annihilator of I in M as a direct product of cyclic groups.

12. In the notation of the preceding exercises prove the following facts about annihilators.
   (a) Let N be a submodule of M and let I be its annihilator in R. Prove that the annihilator of I in M contains N. Give an example where the annihilator of I in M does not equal N.
   (b) Let I be a right ideal of R and let N be its annihilator in M. Prove that the annihilator of N in R contains I. Give an example where the annihilator of N in R does not equal I.

13. Let I be an ideal of R. Let M' be the subset of elements a of M that are annihilated by some power, I^k, of the ideal I, where the power may depend on a (the power of an ideal was defined in Exercise 36 of Section 7.3). Prove that M' is a submodule of M. [Use Exercise 7.]

14. Let z be an element of the center of R, i.e. $zr = rz$ for all $r \in R$. Prove that zM is a submodule of M, where $zM = \{zm \mid m \in M\}$. Show that if R is the ring of 2×2 matrices over a field and e is the matrix with a 1 in position 1,1 and zeros elsewhere then eR is *not* left R-submodule (where $M = R$ is considered as a left R-module as in Example 1) — in this case the matrix e is not in the center of R.

15. If M is a finite abelian group then M is naturally a $\mathbb{Z}$-module. Can this action be extended to make M into a $\mathbb{Q}$-module?

16. Prove that the submodules U_k described in the example of $F[x]$-modules are all of the $F[x]$-submodules for the shift operator.

17. Let T be the shift operator on the vector space V and let $e_1, \ldots, e_n$ be the usual basis vectors described in the example of $F[x]$-modules. If $m \geq n$ find $(a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0)e_n$.

18. Let $F = \mathbb{R}$, let $V = \mathbb{R}^2$ and let T be the linear transformation from V to V which is rotation clockwise about the origin by $\pi/2$ radians. Show that V and 0 are the only $F[x]$-submodules for this T.

19. Let $F = \mathbb{R}$, let $V = \mathbb{R}^2$ and let T be the linear transformation from V to V which is projection onto the y-axis. Show that V, 0, the x-axis and the y-axis are the only $F[x]$-submodules for this T.

20. Let $F = \mathbb{R}$, let $V = \mathbb{R}^2$ and let T be the linear transformation from V to V which is rotation clockwise about the origin by π radians. Show that *every* subspace of V is an $F[x]$-submodule for this T.

21. Let $n \in \mathbb{Z}^+$, $n > 1$ and let R be the ring of $n \times n$ matrices with entries from a field F. Let M be the set of $n \times n$ matrices with arbitrary elements of F in the first column and zeros elsewhere. Show that M is a submodule of R when R is considered as a left module over itself but M is not a submodule of R when R is considered as a right R-module.

10.2 QUOTIENT MODULES AND MODULE HOMOMORPHISMS

This section contains the basic theory of quotient modules and module homomorphisms.

Definition. Let R be a ring and let M and N be R-modules.

(1) A map $\varphi : M \to N$ is an *R-module homomorphism* if it respects the R-module structures of M and N, i.e.
   (a) $\varphi(x + y) = \varphi(x) + \varphi(y)$, for all $x, y \in M$ and
   (b) $\varphi(\alpha x) = \alpha\varphi(x)$, for all $\alpha \in R$, $x \in M$.

(2) An R-module homomorphism is an *isomorphism (of R-modules)* if it is both injective and surjective. The modules M and N are said to be *isomorphic*, denoted $M \cong N$, if there is some R-module isomorphism $\varphi : M \to N$.

(3) If $\varphi : M \to N$ is an R-module homomorphism, let $\ker \varphi = \{m \in M \mid \varphi(m) = 0\}$ (the *kernel* of φ) and let $\varphi(M) = \{n \in N \mid n = \varphi(m) \text{ for some } m \in M\}$ (the *image* of φ, as usual).

(4) Let M and N be R-modules and define $\mathrm{Hom}_R(M, N)$ to be the set of all R-module homomorphisms from M into N.

Any R-module homomorphism is also a homomorphism of the additive groups, but not every group homomorphism need be a module homomorphism (because condition (b) may not be satisfied). The unqualified term "isomorphism" when applied to R-modules will always mean R-module isomorphism. When the symbol $\cong$ is used without qualification it will denote an isomorphism of the respective structures (which will be evident from the context).

It is an easy exercise using the submodule criterion (Proposition 1) to show that kernels and images of R-module homomorphisms are submodules.

Examples

(1) If R is a ring and $M = R$ is a module over itself, then R-module homomorphisms (even from R to itself) need not be ring homomorphisms and ring homomorphisms need not be R-module homomorphisms. For example, when $R = \mathbb{Z}$ the $\mathbb{Z}$-module homomorphism $x \mapsto 2x$ is not a ring homomorphism (1 does not map to 1). When $R = F[x]$ the ring homomorphism $\varphi : f(x) \mapsto f(x^2)$ is not an $F[x]$-module homomorphism (if it were, we would have $x^2 = \varphi(x) = \varphi(x \cdot 1) = x\varphi(1) = x$).

(2) Let R be a ring, let $n \in \mathbb{Z}^+$ and let $M = R^n$. One easily checks that for each $i \in \{1, \ldots, n\}$ the projection map

$$\pi_i : R^n \to R \quad \text{by} \quad \pi_i(x_1, \ldots, x_n) = x_i$$

is a surjective R-module homomorphism with kernel equal to the submodule of n-tuples which have a zero in position i.

(3) If R is a field, R-module homomorphisms are called *linear transformations*. These will be studied extensively in Chapter 11.

(4) For the ring $R = \mathbb{Z}$ the action of ring elements (integers) on any $\mathbb{Z}$-module amounts to just adding and subtracting within the (additive) abelian group structure of the module so that in this case condition (b) of a homomorphism is implied by condition (a). For example, $\varphi(2x) = \varphi(x + x) = \varphi(x) + \varphi(x) = 2\varphi(x)$, etc. It follows that

$\mathbb{Z}$-module homomorphisms are the same as abelian group homomorphisms.

(5) Let R be a ring, let I be a 2-sided ideal of R and suppose M and N are R-modules annihilated by I (i.e. $am = 0$ and $an = 0$ for all $a \in I$, $n \in N$ and $m \in M$). Any R-module homomorphism from N to M is then automatically a homomorphism of (R/I)-modules (see Example 5 of Section 1). In particular, if A is an additive abelian group such that for some prime p, $px = 0$ for all $x \in A$, then any group homomorphism from A to itself is a $\mathbb{Z}/p\mathbb{Z}$-module homomorphism, i.e. is a linear transformation over the field $\mathbb{F}_p$. In particular, the group of all (group) automorphisms of A is the group of invertible linear transformations from A to itself: $GL(A)$.

Proposition 2. Let M, N and L be R-modules.

(1) A map $\varphi : M \to N$ is an R-module homomorphism if and only if $\varphi(\alpha x + y) = \alpha\varphi(x) + \varphi(y)$ for all $x, y \in M$ and all $\alpha \in R$.

(2) Let φ, ψ be elements of $\mathrm{Hom}_R(M, N)$. Define $\varphi + \psi$ by

$$(\varphi + \psi)(m) = \varphi(m) + \psi(m) \qquad \text{for all } m \in M.$$

Then $\varphi + \psi \in \mathrm{Hom}_R(M, N)$ and with this operation $\mathrm{Hom}_R(M, N)$ is an abelian group. If R is a commutative ring then for $\alpha \in R$ define $\alpha\varphi$ by

$$(\alpha\varphi)(m) = \alpha(\varphi(m)) \qquad \text{for all } m \in M.$$

Then $\alpha\varphi \in \mathrm{Hom}_R(M, N)$ and with this action of the commutative ring R the abelian group $\mathrm{Hom}_R(M, N)$ is an R-module.

(3) If $\varphi \in \mathrm{Hom}_R(L, M)$ and $\psi \in \mathrm{Hom}_R(M, N)$ then $\psi \circ \varphi \in \mathrm{Hom}_R(L, N)$.

(4) With addition as above and multiplication defined as function composition, $\mathrm{Hom}_R(M, M)$ is a ring with 1.

Proof: (1) If φ is an R-module homomorphism then certainly $\varphi(\alpha x + y) = \alpha\varphi(x) + \varphi(y)$. Conversely, if $\varphi(\alpha x + y) = \alpha\varphi(x) + \varphi(y)$, take $\alpha = 1$ to see that φ is additive and take $y = 0$ to see that φ commutes with the action of R on M (i.e. is *homogeneous*).

(2) It is straightforward to check that all the abelian group and R-module axioms hold with these definitions — the details are left as an exercise. We note that the commutativity of R is used to show that $\alpha\varphi$ satisfies the second axiom of an R-module

homomorphism, namely:

$$\begin{aligned}(\alpha\varphi)(\beta m) &= \alpha\varphi(\beta m) && \text{(by definition of } \alpha\varphi)\\ &= \alpha\beta(\varphi(m)) && \text{(since } \varphi \text{ is a homomorphism)}\\ &= \beta\alpha\varphi(m) && \text{(since } R \text{ is commutative)}\\ &= \beta(\alpha\varphi)(m) && \text{(by definition of } \alpha\varphi).\end{aligned}$$

Verification of the axioms relies ultimately on the hypothesis that N is an R-module. The domain M could in fact be any set — it does not have to be an R-module nor an abelian group.

(3) Let φ and ψ be as given and let $\alpha \in R$, $x, y \in L$. Then

$$\begin{aligned}(\psi \circ \varphi)(\alpha x + y) &= \psi(\varphi(\alpha x + y))\\ &= \psi(\alpha\varphi(x) + \varphi(y)) && \text{(by (1) applied to } \varphi)\\ &= \alpha\psi(\varphi(x)) + \psi(\varphi(y)) && \text{(by (1) applied to } \psi)\\ &= \alpha(\psi \circ \varphi)(x) + (\psi \circ \varphi)(y)\end{aligned}$$

so, by (1), $\psi \circ \varphi$ is an R-module homomorphism.

(4) Note that since the domain and co-domain of the elements of $\mathrm{Hom}_R(M, M)$ are the same, function composition is defined. By (3), it is a binary operation on $\mathrm{Hom}_R(M, M)$. As usual, function composition is associative. The remaining ring axioms are straightforward to check — the details are left as an exercise. The identity function, I, (as usual, $I(x) = x$, for all $x \in M$) is seen to be the multiplicative identity of $\mathrm{Hom}_R(M, M)$.

Definition. The ring $\mathrm{Hom}_R(M, M)$ is called the *endomorphism ring of* M and will often be denoted by $\mathrm{End}_R(M)$, or just $\mathrm{End}(M)$ when the ring R is clear from the context. Elements of $\mathrm{End}(M)$ are called *endomorphisms.*

When R is commutative there is a natural map from R into $\mathrm{End}(M)$ given by $\alpha \mapsto \alpha I$, where the latter endomorphism of M is just multiplication by α on M (cf. Exercise 7). This ring homomorphism (cf. Exercise 7) from R to $\mathrm{End}_R(M)$ may not be injective since for some α we may have $\alpha m = 0$ for all $m \in M$ (*e.g.* $R = \mathbb{Z}$, $M = \mathbb{Z}/2\mathbb{Z}$, and $\alpha = 2$). When R is a field, however, this map is injective (in general, no unit is in the kernel of this map) and the copy of R in $\mathrm{End}_R(M)$ is called the (subring of) *scalar transformations*.

Next we prove that every submodule N of an R-module M is "normal" in the sense that we can *always* form the quotient module M/N, and the natural projection $\pi : M \to M/N$ is an R-module homomorphism with kernel N. The proof of this fact and, more generally, the subsequent proofs of the isomorphism theorems for modules follow easily from the corresponding facts for groups. The reason for this is because a module is first of all an *abelian* group and so *every* submodule is automatically a normal subgroup and any module homomorphism is, in particular, a homomorphism of abelian groups, all of which we have already considered in Chapter 3. What remains to be proved in order to extend results on abelian groups to corresponding results on modules is to check that the action of R is compatible with these group quotients and

homomorphisms. For example, the map π above was shown to be a group homomorphism in Chapter 3 but the abelian group M/N must be shown to be an R-module (i.e. to have an action by R) and property (b) in the definition of a module homomorphism must be checked for π.

Proposition 3. Let R be a ring, let M be an R-module and let N be a submodule of M. The (additive, abelian) quotient group M/N can be made into an R-module by defining an action of elements of R by

$$\alpha(x + N) = (\alpha x) + N, \qquad \text{for all } \alpha \in R,\ x + N \in M/N.$$

The natural projection map $\pi : M \to M/N$ defined by $\pi(x) = x + N$ is an R-module homomorphism with kernel N.

Proof: Since M is an abelian group under $+$ the quotient group M/N is defined and is an abelian group. To see that the action of the ring element α on the coset $x + N$ is well defined, suppose $x + N = y + N$, i.e. $x - y \in N$. Since N is a (left) R-submodule, $\alpha(x - y) \in N$. Thus $\alpha x - \alpha y \in N$ and $\alpha x + N = \alpha y + N$, as desired. Now since the operations in M/N are "compatible" with those of M, the axioms for an R-module are easily checked in the same way as was done for quotient groups. For example, axiom 2(b) holds as follows: for all $\alpha, \beta \in R$ and $x + N \in M/N$, by definition of the action of ring elements on elements of M/N

$$\begin{aligned}(\alpha\beta)(x + N) &= (\alpha\beta x) + N \\ &= \alpha(\beta x + N) \\ &= \alpha(\beta(x + N)).\end{aligned}$$

The other axioms are similarly checked — the details are left as an exercise. Finally, the natural projection map π described above is, in particular, the natural projection of the abelian group M onto the abelian group M/N hence is a group homomorphism with kernel N. The kernel of any module homomorphism is the same as its kernel when viewed as a homomorphism of the abelian group structures. It remains only to show π is a module homomorphism, i.e. $\pi(\alpha m) = \alpha\pi(m)$. But

$$\begin{aligned}\pi(\alpha m) &= \alpha m + N \\ &= \alpha(m + N) \qquad \text{(by definition of the action of } R \text{ on } M/N) \\ &= \alpha\pi(m).\end{aligned}$$

This completes the proof.

All the isomorphism theorems stated for groups also hold for R-modules. The proofs are similar to that of Proposition 3 above in that they begin by invoking the corresponding theorem for groups and then prove that the group homomorphisms are also R-module homomorphisms. To state the Second Isomorphism Theorem we need the following.

Definition. Let A, B be submodules of the R-module M. The *sum* of A and B is the set

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

One can easily check that the sum of two submodules A and B is a submodule and is the smallest submodule which contains both A and B.

Theorem 4. (Isomorphism Theorems)

(1) (The First Isomorphism Theorem for Modules) Let M, N be R-modules and let $\varphi : M \to N$ be an R-module homomorphism. Then $\ker \varphi$ is a submodule of M and $M/\ker \varphi \cong \varphi(M)$.

(2) (The Second Isomorphism Theorem) Let A, B be submodules of the R-module M. Then $(A + B)/B \cong A/(A \cap B)$.

(3) (The Third Isomorphism Theorem) Let M be an R-module, and let A and B be submodules of M with $A \subseteq B$. Then $(M/A)/(B/A) \cong M/B$.

(4) (The Fourth or Lattice Isomorphism Theorem) Let N be a submodule of the R-module M. There is a bijection between the submodules of M which contain N and the submodules of M/N. The correspondence is given by $A \leftrightarrow (A + N)/N$, for all $A \supseteq N$. This correspondence commutes with the processes of taking sums and intersections (i.e. is a lattice isomorphism between the lattice of submodules of M/N and the lattice of submodules of M which contain N).

Proof: Exercise.

EXERCISES

In these exercises R is a ring with 1 and M is a left R-module.

1. Use the submodule criterion to show that kernels and images of R-module homomorphisms are submodules.
2. Show that the relation "is R-module isomorphic to" is an equivalence relation on any set of R-modules.
3. Give an explicit example of a map from one R-module to another which is a group homomorphism but not an R-module homomorphism.
4. Let A be any $\mathbb{Z}$-module, let a be any element of A and let n be a positive integer. Prove that the map $\varphi_a : \mathbb{Z}/n\mathbb{Z} \to A$ given by $\varphi(\bar{k}) = ka$ is a well-defined $\mathbb{Z}$-module homomorphism if and only if $na = 0$. Prove that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, A) \cong A_n$, where $A_n = \{a \in A \mid na = 0\}$ (so A_n is the annihilator in A of the ideal (n) of $\mathbb{Z}$ — cf. Exercise 10, Section 1).
5. Exhibit all $\mathbb{Z}$-module homomorphisms from $\mathbb{Z}/30\mathbb{Z}$ to $\mathbb{Z}/21\mathbb{Z}$.
6. Prove that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \mathbb{Z}/m\mathbb{Z}) \cong \mathbb{Z}/(n,m)\mathbb{Z}$.
7. Let z be a fixed element of the center of R. Prove that the map $m \mapsto zm$ is an R-module homomorphism from M to itself. Show that for a commutative ring R the map from R to $\mathrm{End}_R(M)$ given by $r \mapsto rI$ is a ring homomorphism (where I is the identity endomorphism).
8. Let $\varphi : M \to N$ be an R-module homomorphism. Prove that $\varphi(\mathrm{Tor}(M)) \subseteq \mathrm{Tor}(N)$ (cf. Exercise 8 in Section 1).
9. Let R be a commutative ring. Prove that $\mathrm{Hom}_R(R, M)$ and M are isomorphic as left R-modules. [Show that each element of $\mathrm{Hom}_R(R, M)$ is determined by its value on the identity of R.]
10. Let R be a commutative ring. Prove that $\mathrm{Hom}_R(R, R)$ and R are isomorphic as rings.

11. Let $A_1, A_2, \ldots, A_n$ be R-modules and let B_i be a submodule of A_i for each $i = 1, 2, \ldots, n$. Prove that

$$(A_1 \times \cdots \times A_n)/(B_1 \times \cdots \times B_n) \cong (A_1/B_1) \times \cdots \times (A_n/B_n).$$

[Recall Exercise 14 in Section 5.1.]

12. Let I be a left ideal of R and let n be a positive integer. Prove

$$R^n/IR^n \cong R/IR \times \cdots \times R/IR \quad (n \text{ times})$$

where IM is defined for any R-module in Exercise 5 of Section 1. [Use the preceding exercise.]

13. Let I be a nilpotent ideal in a commutative ring R (cf. Exercise 37, Section 7.3), let M and N be R-modules and let $\varphi : M \to N$ be an R-module homomorphism. Show that if the induced map $\overline{\varphi} : M/IM \to N/IN$ is surjective, then φ is surjective.

10.3 GENERATION OF MODULES, DIRECT SUMS AND FREE MODULES

Let R be a ring. As in the preceding sections the term "module" will mean "left module." We first extend the notion of the sum of two submodules to sums of any finite number of submodules and define the submodule generated by a subset.

Definition. Let M be an R-module and let $N_1, \ldots, N_n$ be submodules of M.

(1) The *sum* of $N_1, \ldots, N_n$ is the set of all finite sums of elements from the sets N_i: $\{a_1 + a_2 + \cdots + a_n \mid a_i \in N_i \text{ for all } i\}$. Denote this sum by $N_1 + \cdots + N_n$.

(2) For any subset A of M let

$$RA = \{\alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_m a_m \mid \alpha_1, \ldots, \alpha_m \in R,\ a_1, \ldots, a_m \in A,\ m \in \mathbb{Z}^+\}$$

(where the convention is $RA = \{0\}$ if $A = \emptyset$). If A is the finite set $\{a_1, a_2, \ldots, a_n\}$ we shall write $Ra_1 + Ra_2 + \cdots + Ra_n$ for RA. Call RA the *submodule of M generated by A*. If N is a submodule of M (possibly $N = M$) and $N = RA$, for some subset A of M, we call A a *set of generators* or *generating set* for N, and we say *N is generated by A*.

(3) A submodule N of M (possibly $N = M$) is *finitely generated* if there is some finite subset A of M such that $N = RA$, that is, if N is generated by some finite subset.

(4) A submodule N of M (possibly $N = M$) is *cyclic* if there exists an element $a \in M$ such that $N = Ra$, that is, if N is generated by one element:

$$N = Ra = \{ra \mid r \in R\}.$$

It is easy to see using the Submodule Criterion that for any subset A of M, RA is indeed a submodule of M and is the smallest submodule of M which contains A (i.e. any submodule of M which contains A also contains RA). In particular, for submodules $N_1, \ldots, N_n$ of M, $N_1 + \cdots + N_n$ is just the submodule generated by the set $N_1 \cup \cdots \cup N_n$ and is the smallest submodule of M containing N_i, for all i. If $N_1, \ldots, N_n$ are generated by sets $A_1, \ldots, A_n$ respectively, then $N_1 + \cdots + N_n$ is generated by $A_1 \cup \cdots \cup A_n$. Note that cyclic modules are, a fortiori, finitely generated.

A submodule N of an R-module M may have many different generating sets (for instance the set N itself always generates N). If N is finitely generated, then there is a smallest nonnegative integer d such that N is generated by d elements (and no fewer). Any generating set consisting of d elements will be called a *minimal set of generators for* N (it is not unique in general). If N is not finitely generated, it need not have a minimal generating set.

The process of generating submodules of an R-module M by taking subsets A of M and forming all finite "R-linear combinations" of elements of A will be our primary way of producing submodules (this notion is perhaps familiar from vector space theory where it is referred to as taking the *span* of A). The obstruction which made the analogous process so difficult for groups in general was the noncommutativity of group operations. For abelian groups, G, however, it was much simpler to control the subgroup $\langle A \rangle$ generated by A, for a subset A of G (see Section 2.4 for the complete discussion of this). The situation for R-modules is similar to that of abelian groups (even if R is a noncommutative ring) because we can always collect "like terms" in elements of A, i.e. terms such as $\alpha_1 a_1 + \alpha_2 a_2 + \beta_1 a_1$ can always be simplified to $(\alpha_1 + \beta_1)a_1 + \alpha_2 a_2$. This again reflects the underlying abelian group structure of modules.

Examples

(1) Let R $= \mathbb{Z}$ and let M be any R-module, that is, any abelian group. If $a \in M$, then $\mathbb{Z}a$ is just the cyclic subgroup of M generated by a: $\langle a \rangle$ (compare Definition 4 above with the definition of a cyclic group). More generally, M is generated as a $\mathbb{Z}$-module by a set A if and only if M is generated as a group by A (that is, the action of ring elements in this instance produces no elements that cannot already be obtained from A by addition and subtraction). The definition of finitely generated for $\mathbb{Z}$-modules is identical to that for abelian groups found in Chapter 5.

(2) Let R be any ring with 1 and let M be the (left) R-module R itself. Note that R is a finitely generated, in fact cyclic, R-module because $R = R1$ (i.e. we can take $A = \{1\}$). Recall that the submodules of R are precisely the left ideals of R, so saying I is a cyclic R-submodule of the left R-module R is the same as saying I is a principal ideal of R (usually the term "principal ideal" is used in the context of commutative rings). Also, saying I is a finitely generated R-submodule of R is the same as saying I is a finitely generated ideal. When R is a commutative ring we often write AR or aR for the submodule (ideal) generated by A or a respectively, as we have been doing for $\mathbb{Z}$ when we wrote $n\mathbb{Z}$. In this situation $AR = RA$ and $aR = Ra$ (elementwise as well). Thus a Principal Ideal Domain is a (commutative) integral domain R with identity in which every R-submodule of R is cyclic.

Submodules of a finitely generated module need not be finitely generated: take M to be the cyclic R-module R itself where R is the polynomial ring in infinitely many variables $x_1, x_2, x_3, \ldots$ with coefficients in some field F. The submodule (i.e. 2-sided ideal) generated by $\{x_1, x_2, \ldots\}$ cannot be generated by any finite set (note that one must show that *no* finite subset of this ideal will generate it).

(3) Let R be any ring and let M be the free module of rank n over R, as described in the first section. For each $i \in \{1, 2, \ldots, n\}$ let $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 appears in position i. Since

$$(\alpha_1, \alpha_2, \ldots, \alpha_n) = \sum_{i=i}^{n} \alpha_i e_i$$

it is clear that M is generated by $\{e_1, \ldots, e_n\}$. It is more difficult to show that this is a *minimal* generating set (for commutative rings this follows from Exercise 2).

(4) Let F be a field, let x be an indeterminate, let V be a vector space over F and let T be a linear transformation from V to V. Make V into an $F[x]$-module via T. Then V is a *cyclic* $F[x]$-module (with generator v) if and only if $V = \{p(x)v \mid p(x) \in F[x]\}$, that is, if and only if every element of V can be written as an F-linear combination of elements of the set $\{T^n(v) \mid n \geq 0\}$. This in turn is equivalent to saying $\{v, T(v), T^2(v), \ldots\}$ span V as a vector space over F.

For instance if T is the identity linear transformation from V to V or the zero linear transformation, then for every $v \in V$ and every $p(x) \in F[x]$ we have $p(x)v = \alpha v$ for some $\alpha \in F$. Thus if V has dimension > 1, V cannot be a cyclic $F[x]$-module.

For another example suppose V is affine n-space and T is the "shift operator" described in Section 1. Let e_i be the ith basis vector (as usual) numbered so that T is defined by $T^k(e_n) = e_{n-k}$ for $1 \leq k < n$. Thus V is spanned by the elements $e_n, T(e_n), \ldots, T^{n-1}(e_n)$, that is, V is a cyclic $F[x]$-module with generator e_n. For $n > 1$, V is not, however, a cyclic F-module (i.e. is not a 1-dimensional vector space over F).

Definition. Let $M_1, \ldots, M_k$ be a collection of R-modules. The collection of n-tuples $(m_1, m_2, \ldots, m_k)$ where $m_i \in M_i$ with addition and action of R defined componentwise is called the *direct product* of $M_1, \ldots, M_k$, denoted $M_1 \times \cdots \times M_k$.

It is evident that the direct product of a collection of R-modules is again an R-module. The direct product of $M_1, \ldots, M_k$ is also referred to as the *(external) direct sum* of $M_1, \ldots, M_k$ and denoted $M_1 \oplus \cdots \oplus M_k$. The direct product and direct sum of an infinite number of modules (which are different in general) are defined in Exercise 20.

The next proposition indicates when a module is isomorphic to the direct product of some of its submodules and is the analogue for modules of Theorem 9 in Section 5.4 (which determines when a group is the direct product of two of its subgroups).

Proposition 5. Let $N_1, N_2, \ldots, N_k$ be submodules of the R-module M. Then the following are equivalent:

(1) The map $\pi : N_1 \times N_2 \times \cdots \times N_k \to N_1 + N_2 + \cdots + N_k$ defined by

$$\pi(a_1, a_2, \ldots, a_k) = a_1 + a_2 + \cdots + a_k$$

is an isomorphism (of R-modules): $N_1 + N_2 + \cdots + N_k \cong N_1 \times N_2 \times \cdots \times N_k$

(2) $N_j \cap (N_1 + N_2 + \cdots + N_{j-1} + N_{j+1} + \cdots + N_k) = 0$ for all $j \in \{1, 2, \ldots, k\}$

(3) Every $x \in N_1 + \cdots + N_k$ can be written *uniquely* in the form $a_1 + a_2 + \cdots + a_k$ with $a_i \in N_i$.

Proof: To prove (1) implies (2), suppose for some j that (2) fails to hold and let $a_j \in (N_1 + \cdots + N_{j-1} + N_{j+1} + \cdots + N_k) \cap N_j$, with $a_j \neq 0$. Then

$$a_j = a_1 + \cdots + a_{j-1} + a_{j+1} + \cdots + a_k$$

for some $a_i \in N_i$, and $(a_1, \ldots, a_{j-1}, -a_j, a_{j+1}, \ldots, a_k)$ would be a nonzero element of $\ker \pi$, a contradiction.

Assume now that (2) holds. We prove by induction that for each j every element of $N_1 + N_2 + \cdots + N_j$ can be written uniquely as $a_1 + a_2 + \cdots + a_j$, where $a_i \in N_i$ for all i. Applying this result to $j = k$ will show that (2) implies (3).

The induction assertion is clearly true for $j = 0$ or 1. Assume the result is true for $j - 1$. Let $x \in N_1 + N_2 + \cdots + N_j$. Suppose that x can be written in two distinct ways:

$$x = a_1 + a_2 + \cdots + a_j = b_1 + b_2 + \cdots + b_j \qquad a_i, b_i \in N_i,\ 1 \le i \le j.$$

Subtracting the two equal sums and collecting the jth terms on the left gives

$$\begin{aligned} a_j - b_j &= (b_1 + b_2 + \cdots + b_{j-1}) - (a_1 + a_2 + \cdots + a_{j-1}) \\ &\in N_j \cap (N_1 + N_2 + \cdots + N_{j-1}) \end{aligned}$$

Since we are assuming (2) we must have $a_j - b_j = 0$, i.e. $a_j = b_j$ and

$$a_1 + a_2 + \cdots + a_{j-1} = b_1 + b_2 + \ldots b_{j-1}.$$

By our induction assumption we must have $a_i = b_i$ for $1 \le i \le j - 1$, so that $a_i = b_i$ for all i. This completes the induction and establishes (3).

Finally, to see that (3) implies (1) observe first that the map π is clearly a surjective R-module homomorphism. Then (3) simply implies π is injective, hence is an isomorphism, completing the proof.

If an R-module $M = N_1 + N_2 + \ldots + N_k$ is the sum of submodules $N_1, N_2, \ldots, N_k$ of M satisfying the equivalent conditions of the proposition above, then M is said to be the *(internal) direct sum* of $N_1, N_2, \ldots, N_k$, written

$$M = N_1 \oplus N_2 \oplus \cdots \oplus N_k.$$

By the proposition, this is equivalent to the assertion that every element m of M can be written *uniquely* as a sum of elements $m = n_1 + n_2 + \ldots + n_k$ with $n_i \in N_i$. (Note that part (1) of the proposition is the statement that the internal direct sum of $N_1, N_2, \ldots, N_k$ is isomorphic to their external direct sum, which is the reason we identify them and use the same notation for both.)

Definition. An R-module F is said to be *free* on the subset A of F if for every element x of F, there exist unique elements $\alpha_1, \alpha_2, \ldots, \alpha_n$ of R and unique $a_1, a_2, \ldots, a_n$ in A such that $x = \alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_n a_n$, for some $n \in \mathbb{Z}^+$. In this situation we say A is a *basis* or *set of free generators* for F and the cardinality of A is called the *rank* of F.

One should be careful to note the difference between the uniqueness property of direct sums (Proposition 5(3)) and the uniqueness property of free modules. Namely, in the direct sum of two modules, say $N_1 \oplus N_2$, each element can be written uniquely as $n_1 + n_2$; here the uniqueness refers to the *module elements* n_1 and n_2. In the case of free modules, the uniqueness is on the *ring elements as well as the module elements*. For example, if $R = \mathbb{Z}$ and $N_1 = N_2 = \mathbb{Z}/2\mathbb{Z}$, then each element of $N_1 \oplus N_2$ has

a unique representation in the form $n_1 + n_2$ where each $n_i \in N_i$, however n_1 (for instance) can be expressed as n_1 or $3n_1$ or $5n_1 \ldots$ etc., so each element does not have a unique representation in the form $\alpha_1 a_1 + \alpha_2 a_2$, where $\alpha_1, \alpha_2 \in R$, $a_1 \in N_1$ and $a_2 \in N_2$. Thus $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ is not a free $\mathbb{Z}$-module.

Theorem 6. For any set A there is a free R-module $F(A)$ on the set A and $F(A)$ satisfies the following *universal property:* if M is any R-module and $\varphi : A \to M$ is any map of sets, then there is a unique R-module homomorphism $\Phi : F(A) \to M$ such that $\Phi(a) = \varphi(a)$, for all $a \in A$, that is, the following diagram commutes:

$$\begin{tikzcd} A \arrow[r, "\text{inclusion}"] \arrow[dr, "\varphi"'] & F(A) \arrow[d, "\Phi"] \\ & M \end{tikzcd}$$

When A is the finite set $\{a_1, a_2, \ldots, a_n\}$, $F(A) = Ra_1 \oplus Ra_2 \oplus \cdots \oplus Ra_n \cong R^n$. (Compare: Section 6.3, free groups.)

Proof: Let $F(A) = \{0\}$ if $A = \emptyset$. If A is nonempty let $F(A)$ be the collection of all set functions $f : A \to R$ such that $f(a) = 0$ for all but finitely many $a \in A$. Make $F(A)$ into an R-module by pointwise addition of functions and pointwise multiplication of a ring element times a function, i.e.

$$\begin{aligned} (f+g)(a) &= f(a) + g(a) \quad \text{and} \\ (\alpha f)(a) &= \alpha(f(a)), \qquad \text{for all } a \in A, \alpha \in R \text{ and } f, g \in F(A). \end{aligned}$$

It is an easy matter to check that all the R-module axioms hold (the details are omitted). Identify A as a subset of $F(A)$ by $a \mapsto f_a$, where f_a is the function which is 1 at a and zero elsewhere. We can, in this way, think of $F(A)$ as all finite R-linear combinations of elements of A by identifying each function f with the sum $\alpha_1 a_1 + \alpha_2 a_2 + \cdots + \alpha_n a_n$, where f takes on the value α_i at a_i and is zero at all other elements of A. Moreover, each element of $F(A)$ has a unique expression as such a formal sum. To establish the universal property of $F(A)$ suppose $\varphi : A \to M$ is a map of the set A into the R-module M. Define $\Phi : F(A) \to M$ by

$$\Phi : \sum_{i=1}^{n} \alpha_i a_i \mapsto \sum_{i=1}^{n} \alpha_i \varphi(a_i).$$

By the uniqueness of the expression for the elements of $F(A)$ as linear combinations of the a_i we see easily that Φ is a well defined R-module homomorphism (the details are left as an exercise). By definition, the restriction of Φ to A equals φ. Finally, since $F(A)$ is generated by A, once we know the values of an R-module homomorphism on A its values on every element of $F(A)$ are uniquely determined, so Φ is the unique extension of φ to all of $F(A)$.

When A is the finite set $\{a_1, a_2, \ldots, a_n\}$ Proposition 5(3) shows that $F(A) = Ra_1 \oplus Ra_2 \oplus \cdots \oplus Ra_n$. Since $R \cong Ra_i$ for all i (under the map $r \mapsto ra_i$) Proposition 5(1) shows that the direct sum is isomorphic to R^n.

Corollary 7

(1) If F_1 and F_2 are free modules on the same set A, there is a unique isomorphism between F_1 and F_2 which is the identity map on A.

(2) If F is any free R-module with basis A, then $F \cong F(A)$. In particular, F enjoys the same universal property with respect to the set A as $F(A)$ does in Theorem 6.

Proof: Exercise.

If F is a free R-module with basis A, we shall often (particularly in the case of vector spaces) define R-module homomorphisms from F into other R-modules simply by specifying their values on the elements of A and then saying *"extend by linearity"*. Corollary 7(2) ensures that this is permissible.

When $R = \mathbb{Z}$, the free module on a set A is called the *free abelian group on* A. If $|A| = n$, $F(A)$ is called the free abelian group of *rank* n and is isomorphic to $\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}$ (n times). These definitions agree with the ones given in Chapter 5.

EXERCISES

In these exercises R is a ring with 1 and M is a left R-module.

1. Prove that if A and B are sets of the same cardinality, then the free modules $F(A)$ and $F(B)$ are isomorphic.
2. Assume R is commutative. Prove that $R^n \cong R^m$ if and only if $n = m$, i.e. two free R-modules of finite rank are isomorphic if and only if they have the same rank. [Apply Exercise 12 of Section 2 with I a maximal ideal of R. You may assume that if F is a field, then $F^n \cong F^m$ if and only if $n = m$, i.e. two finite dimensional vector spaces over F are isomorphic if and only if they have the same dimension — this will be proved in Section 11.1.]
3. Show that the $F[x]$-modules in Exercises 18 and 19 of Section 1 are both cyclic.
4. An R-module M is called a *torsion* module if for each $m \in M$ there is a nonzero element $r \in R$ such that $rm = 0$, where r may depend on m (i.e. $M = \mathrm{Tor}(M)$ in the notation of Exercise 8 of Section 1). Prove that every finite abelian group is a torsion $\mathbb{Z}$-module. Give an example of an infinite abelian group that is a torsion $\mathbb{Z}$-module.
5. Let R be an integral domain. Prove that every finitely generated torsion R-module has a nonzero annihilator i.e. there is a nonzero element $r \in R$ such that $rm = 0$ for all $m \in M$ — here r does not depend on m (the annihilator of a module was defined in Exercise 9 of Section 1). Give an example of a torsion R-module whose annihilator is the zero ideal.
6. Prove that if M is a finitely generated R-module that is generated by n elements then every quotient of M may be generated by n (or fewer) elements. Deduce that quotients of cyclic modules are cyclic.
7. Let N be a submodule of M. Prove that if both M/N and N are finitely generated then so is M.
8. Let S be the collection of sequences $(a_1, a_2, a_3, \ldots)$ of integers $a_1, a_2, a_3, \ldots$ where all but finitely many of the a_i are 0 (called the *direct sum* of infinitely many copies of $\mathbb{Z}$). Recall that S is a ring under componentwise addition and multiplication and S does not have a multiplicative identity — cf. Exercise 20, Section 7.1. Prove that S is not finitely generated as a module over itself.

9. An R-module M is called *irreducible* if $M \neq 0$ and if 0 and M are the only submodules of M. Show that M is irreducible if and only if $M \neq 0$ and M is a cyclic module with any nonzero element as generator. Determine all the irreducible $\mathbb{Z}$-modules.

10. Assume R is commutative. Show that an R-module M is irreducible if and only if M is isomorphic (as an R-module) to R/I where I is a maximal ideal of R. [By the previous exercise, if M is irreducible there is a natural map $R \to M$ defined by $r \mapsto rm$, where m is any fixed nonzero element of M.]

11. Show that if M_1 and M_2 are irreducible R-modules, then any nonzero R-module homomorphism from M_1 to M_2 is an isomorphism. Deduce that if M is irreducible then $\mathrm{End}_R(M)$ is a division ring (this result is called *Schur's Lemma*). [Consider the kernel and the image.]

12. Let R be a commutative ring and let A, B and M be R-modules. Prove the following isomorphisms of R-modules:
    (a) $\mathrm{Hom}_R(A \times B, M) \cong \mathrm{Hom}_R(A, M) \times \mathrm{Hom}_R(B, M)$
    (b) $\mathrm{Hom}_R(M, A \times B) \cong \mathrm{Hom}_R(M, A) \times \mathrm{Hom}_R(M, B)$.

13. Let R be a commutative ring and let F be a free R-module of finite rank. Prove the following isomorphism of R-modules: $\mathrm{Hom}_R(F, R) \cong F$.

14. Let R be a commutative ring and let F be the free R-module of rank n. Prove that $\mathrm{Hom}_R(F, M) \cong M \times \cdots \times M$ (n times). [Use Exercise 9 in Section 2 and Exercise 12.]

15. An element $e \in R$ is called a *central idempotent* if $e^2 = e$ and $er = re$ for all $r \in R$. If e is a central idempotent in R, prove that $M = eM \oplus (1-e)M$. [Recall Exercise 14 in Section 1.]

The next two exercises establish the Chinese Remainder Theorem for modules (cf. Section 7.6).

16. For any ideal I of R let IM be the submodule of M defined in Exercise 5 of Section 1. Let $A_1, \ldots, A_k$ be any ideals in the ring R. Prove that the map

$$M \to M/A_1M \times \cdots \times M/A_kM \qquad \text{defined by} \qquad m \mapsto (m + A_1M, \ldots, m + A_kM)$$

is an R-module homomorphism with kernel $A_1M \cap A_2M \cap \cdots \cap A_kM$.

17. Continuing in the notation of the preceding exercise, assume further that the ideals $A_1, \ldots, A_k$ are pairwise comaximal (i.e. $A_i + A_j = R$ for all $i \neq j$). Prove that

$$M/(A_1 \cdots A_k)M \cong M/A_1M \times \cdots \times M/A_kM.$$

[See the proof of the Chinese Remainder Theorem for rings in Section 7.6.]

18. Let R be a Principal Ideal Domain and let M be an R-module that is annihilated by the nonzero, proper ideal (a). Let $a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ be the unique factorization of a into distinct prime powers in R. Let M_i be the annihilator of $p_i^{\alpha_i}$ in M, i.e. $M_i = \{m \in M \mid p_i^{\alpha_i} m = 0\}$ — called the *p_i-primary component of M*. Prove that

$$M = M_1 \oplus M_2 \oplus \cdots \oplus M_k.$$

19. Show that if M is a finite abelian group of order $a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ then, considered as a $\mathbb{Z}$-module, M is annihilated by (a), the p_i-primary component of M is the unique Sylow p_i-subgroup of M and M is isomorphic to the direct product of its Sylow subgroups.

20. Let I be a nonempty index set and for each $i \in I$ let M_i be an R-module. The *direct product* of the modules M_i is defined to be their direct product as abelian groups (cf. Exercise 15 in Section 5.1) with the action of R componentwise multiplication. The *direct sum* of the modules M_i is defined to be the restricted direct product of the abelian groups M_i (cf. Exercise 17 in Section 5.1) with the action of R componentwise multiplication. In other words,

the direct sum of the M_i's is the subset of the direct product, $\prod_{i \in I} M_i$, which consists of all elements $\prod_{i \in I} m_i$ such that only finitely many of the components m_i are nonzero; the action of R on the direct product or direct sum is given by $r \prod_{i \in I} m_i = \prod_{i \in I} r m_i$ (cf. the Appendix for the definition of Cartesian products of infinitely many sets).

(a) Prove that the direct product of the M_i's is an R-module and the direct sum of the M_i's is a submodule of their direct product.

(b) Show that if $R = \mathbb{Z}$, $I = \mathbb{Z}^+$ and M_i is the cyclic group of order i for each i, then the direct sum of the M_i's is not isomorphic to their direct product. [Look at torsion.]

21. Let I be a nonempty index set and for each $i \in I$ let N_i be a submodule of M. Prove that the following are equivalent:

(i) the submodule of M generated by all the N_i's is isomorphic to the direct sum of the N_i's

(ii) if $\{i_1, i_2, \ldots, i_k\}$ is any finite subset of I then $N_{i_1} \cap (N_{i_2} + \cdots + N_{i_k}) = 0$

(iii) if $\{i_1, i_2, \ldots, i_k\}$ is any finite subset of I then $N_1 + \cdots + N_k = N_1 \oplus \cdots \oplus N_k$

(iv) for every element x of the submodule of M generated by the N_i's there are unique elements $a_i \in N_i$ for all $i \in I$ such that all but a finite number of the a_i are zero and x is the (finite) sum of the a_i.

22. Let R be a Principal Ideal Domain, let M be a torsion R-module (cf. Exercise 4) and let p be a prime in R (do not assume M is finitely generated, hence it need not have a nonzero annihilator — cf. Exercise 5). The *p-primary component* of M is the set of all elements of M that are annihilated by some positive power of p.

(a) Prove that the p-primary component is a submodule. [See Exercise 13 in Section 1.]

(b) Prove that this definition of p-primary component agrees with the one given in Exercise 18 when M has a nonzero anihilator.

(c) Prove that M is the (possibly infinite) direct sum of its p-primary components, as p runs over all primes of R.

11

Vector Spaces

In this chapter we review the basic theory of finite dimensional vector spaces over an arbitrary field F (some infinite dimensional vector space theory is covered in the exercises). Since the proofs are identical to the corresponding arguments for real vector spaces our treatment is very terse. For the most part we include only those results which are used in other parts of the text so basic topics such as Gauss–Jordan elimination, row echelon forms, methods for finding bases of subspaces, elementary properties of matrices, etc. are not covered or are discussed in the exercises. The reader should therefore consider this chapter as a refresher in linear algebra and as a prelude to field theory and Galois theory. Characteristic polynomials and eigenvalues will be reviewed and treated in a larger context in the next chapter.

11.1 DEFINITIONS AND BASIC THEORY

The terminology for vector spaces is slightly different from that of modules, that is, when the ring R is a field there are different names for many of the properties of R-modules which we defined in the last chapter. The following is a dictionary of these new terms (many of which may already be familiar). The definition of each corresponding vector space property is the same (verbatim) as the module-theoretic definition with the only added assumption being that the ring R is a field (so these definitions are not repeated here).

Terminology for R any Ring	Terminology for R a Field
M is an R-module	M is a vector space over R
m is an element of M	m is a vector in M
a is a ring element	a is a scalar
N is a submodule of M	N is a subspace of M
M/N is a quotient module	M/N is a quotient space
M is a free module of rank n	M is a vector space of dimension n
M is a finitely generated module	M is a finite dimensional vector space
M is a nonzero cyclic module	M is a 1-dimensional vector space
$\varphi : M \to N$ is an R-module homomorphism	$\varphi : M \to N$ is a linear transformation
M and N are isomorphic R-modules	M and N are isomorphic vector spaces
the subset A of M generates M	the subset A of M spans M
$M = RA$	each element of M is a linear combination of elements of A i.e. $M = \text{Span}(A)$

For the remainder of this chapter F is a field and V is a vector space over F.

One of the first results we shall prove about vector spaces is that they are free F-modules, that is, they have bases. Although our arguments treat only the case of finite dimensional spaces, the corresponding result for arbitrary vector spaces is proved in the exercises as an application of Zorn's Lemma. The reader may first wish to review the section in the previous chapter on free modules, especially their properties pertaining to homomorphisms.

Definition

(1) A subset S of V is called a set of *linearly independent* vectors if $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$ with $\alpha_1, \alpha_2, \ldots, \alpha_n \in F$ and $v_1, v_2, \ldots, v_n \in S$ implies $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$.

(2) A *basis* of a vector space V is an ordered set of linearly independent vectors which span V. In particular two bases will be considered different even if one is simply a rearrangement of the other. This is sometimes referred to as an *ordered basis*.

Examples

(1) The space $V = F[x]$ of polynomials in the variable x with coefficients from the field F is in particular a vector space over F. The elements $1, x, x^2, \ldots$ are linearly independent by definition (i.e. a polynomial is 0 if and only if all its coefficients are 0). Since these elements also span V by definition, they are a basis for V.

(2) The collection of solutions of a linear, homogeneous, constant coefficient differential equation (for example, $y'' - 3y' + 2y = 0$) over $\mathbb{C}$ form a vector space over $\mathbb{C}$ since differentiation is a linear operator. Elements of this vector space are linearly independent if they are linearly independent as functions. For example, e^t and e^{2t} are easily seen to be solutions of the equation $y'' - 3y' + 2y = 0$ (differentiation with respect to t). They are linearly independent functions since $ae^t + be^{2t} = 0$ implies $a + b = 0$ (let $t = 0$) and $ae + be^2 = 0$ (let $t = 1$) and the only solution to these two equations is $a = b = 0$. It is a theorem in differential equations that these elements span the set of solutions of this equation, hence are a basis for this space.

Proposition 1. Assume the set $\mathcal{A} = \{v_1, v_2, \ldots, v_n\}$ spans the vector space V but no proper subset of $\mathcal{A}$ spans V. Then $\mathcal{A}$ is a basis of V. In particular, any finitely generated (i.e. finitely spanned) vector space over F is a free F-module.

Proof: It is only necessary to prove that $v_1, v_2, \ldots, v_n$ are linearly independent. Suppose $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$ where not all of the α_i are 0. By reordering, we may assume that $\alpha_1 \neq 0$ and then

$$v_1 = -\frac{1}{\alpha_1}(\alpha_2 v_2 + \cdots + \alpha_n v_n).$$

It follows that $\{v_2, v_3, \ldots, v_n\}$ also spans V since any linear combination of $v_1, v_2, \ldots, v_n$ can be written as a linear combination of $v_2, v_3, \ldots, v_n$ using the equation above. This is a contradiction.

Example

Let F be a field and consider $F[x]/(f(x))$ where $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. The ideal $(f(x))$ is a subspace of the vector space $F[x]$ and the quotient $F[x]/(f(x))$ is also a vector space over F. By the Euclidean Algorithm, every polynomial $a(x) \in F[x]$ can be written uniquely in the form $a(x) = q(x)f(x) + r(x)$ where $r(x) \in F[x]$ and $0 \leq \deg r(x) \leq n-1$. Since $q(x)f(x) \in (f(x))$, it follows that every element of the quotient is represented by a polynomial $r(x)$ of degree $\leq n-1$. Two distinct such polynomials cannot be the same in the quotient since this would say their difference (which is a nonzero polynomial of degree at most $n-1$) would be divisible by $f(x)$ (which is of degree n). It follows that the elements $\bar{1}, \bar{x}, \overline{x^2}, \ldots, \overline{x^{n-1}}$ (the bar denotes the image of these elements in the quotient, as usual) span $F[x]/(f(x))$ as a vector space over F and that no proper subset of these elements also spans, hence these elements give a basis for $F[x]/(f(x))$.

Corollary 2. Assume the finite set $\mathcal{A}$ spans the vector space V. Then $\mathcal{A}$ contains a basis of V.

Proof: Any subset $\mathcal{B}$ of $\mathcal{A}$ spanning V such that no proper subset of $\mathcal{B}$ also spans V (there clearly exist such subsets) is a basis for V by Proposition 1.

Theorem 3. (A Replacement Theorem) Assume $\mathcal{A}$ is a basis for V containing n elements and $\{b_1, b_2, \ldots, b_m\}$ is a set of linearly independent vectors in V. Then there is an ordering $a_1, a_2, \ldots, a_n$ such that for each $k \in \{1, 2, \ldots, m\}$ the set $\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$ is a basis of V. In other words, the elements $b_1, b_2, \ldots, b_m$ can be used to successively replace the elements of the basis $\mathcal{A}$, still retaining a basis. In particular, $n \geq m$.

Proof: Proceed by induction on k. If $k = 0$ there is nothing to prove, since $\mathcal{A}$ is given as a basis for V. Suppose now that $\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$ is a basis for V. Then in particular this is a spanning set, so b_{k+1} is a linear combination:

$$b_{k+1} = \beta_1 b_1 + \cdots + \beta_k b_k + \alpha_{k+1}a_{k+1} + \cdots + \alpha_n a_n. \tag{11.1}$$

Not all of the α_i can be 0, since this would imply b_{k+1} is a linear combination of $b_1, b_2, \ldots, b_k$, contrary to the linear independence of these elements. By reordering if necessary, we may assume $\alpha_{k+1} \neq 0$. Then solving this last equation for a_{k+1} as a linear combination of b_{k+1} and $b_1, b_2, \ldots, b_k, a_{k+2}, \ldots, a_n$ shows

$$\text{Span}\{b_1, b_2, \ldots, b_k, b_{k+1}, a_{k+2}, \ldots, a_n\} = \text{Span}\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\} = V$$

and so this is a spanning set for V. It remains to see that $\{b_1, b_2, \ldots, b_k, b_{k+1}, a_{k+2}, \ldots, a_n\}$ are linearly independent. If

$$\beta_1 b_1 + \cdots + \beta_k b_k + \beta_{k+1}b_{k+1} + \alpha_{k+2}a_{k+2} + \cdots + \alpha_n a_n = 0 \tag{11.2}$$

then substituting for b_{k+1} from the expression for b_{k+1} in equation (1), we obtain a linear combination of $\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$ equal to 0, where the coefficient of a_{k+1} is β_{k+1}. Since this last set is a basis by induction, all the coefficients in

this linear combination, in particular β_{k+1}, must be 0. But then equation (2) is

$$\beta_1 b_1 + \cdots + \beta_k b_k + \alpha_{k+2} a_{k+2} + \cdots + \alpha_n a_n = 0.$$

Again by the induction hypothesis all the other coefficients must be 0 as well. Thus $\{b_1, b_2, \ldots, b_k, b_{k+1}, a_{k+2}, \ldots, a_n\}$ is a basis for V, and the induction is complete.

Corollary 4

(1) Suppose V has a finite basis with n elements. Any set of linearly independent vectors has $\leq n$ elements. Any spanning set has $\geq n$ elements.

(2) If V has some finite basis then any two bases of V have the same cardinality.

Proof: (1) This is a restatement of the last result of Theorem 3 and Corollary 2. (2) This is immediate from (1) since a basis is both a spanning set and a linearly independent set.

Definition. If V is a finitely generated F-module (i.e. has a finite basis) the cardinality of any basis is called the *dimension* of V and is denoted by $\dim_F V$, or just $\dim V$ when F is clear from the context, and V is said to be *finite dimensional* over F. If V is not finitely generated, V is said to be infinite dimensional (written $\dim V = \infty$).

Examples

(1) The dimension of the space of solutions to the differential equation $y'' - 3y' + 2y = 0$ over $\mathbb{C}$ is 2 (with basis e^t, e^{2t}, for example). In general, it is a theorem in differential equations that the space of solutions of an nth order, linear, homogeneous, constant coefficient differential equation of degree n over $\mathbb{C}$ form a vector space over $\mathbb{C}$ of dimension n.

(2) The dimension over F of the quotient $F[x]/(f(x))$ by the nonzero polynomial $f(x)$ considered above is $n = \deg f(x)$. The space $F[x]$ and its subspace $(f(x))$ are infinite dimensional vector spaces over F.

Corollary 5. (Building-Up Lemma) If A is a set of linearly independent vectors in the finite dimensional space V then there exists a basis of V containing A.

Proof: This is also immediate from Theorem 3, since we can use the elements of A to successively replace the elements of any given basis for V (which exists by the assumption that V is finite dimensional).

Theorem 6. If V is an n dimensional vector space over F, then $V \cong F^n$. In particular, any two finite dimensional vector spaces over F of the same dimension are isomorphic.

Proof: Let $v_1, v_2, \ldots, v_n$ be a basis for V. Define the map

$$\varphi : F^n \to V \qquad \text{by} \qquad \varphi(\alpha_1, \alpha_2, \ldots, \alpha_n) = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n.$$

The map φ is clearly F-linear, is *surjective* since the v_i *span* V, and is *injective* since the v_i are *linearly independent,* hence is an isomorphism.

Examples

(1) Let $\mathbb{F}$ be a finite field with q elements and let W be a k-dimensional vector space over $\mathbb{F}$. We show that the number of distinct bases of W is

$$(q^k - 1)(q^k - q)(q^k - q^2)\dots(q^k - q^{k-1}).$$

Every basis of W can be built up as follows. Any nonzero vector w_1 can be the first element of a basis. Since W is isomorphic to $\mathbb{F}^k$, $|W| = q^k$, so there are $q^k - 1$ choices for w_1. Any vector not in the 1-dimensional space spanned by w_1 is linearly independent from w_1 and so may be chosen for the second basis element, w_2. A 1-dimensional space is isomorphic to $\mathbb{F}$ and so has q elements. Thus there are $q^k - q$ choices for w_2. Proceeding in this way one sees that at the ith stage any vector not in the $(i-1)$-dimensional space spanned by $w_1, w_2, \dots, w_{i-1}$ will be linearly independent from $w_1, w_2, \dots, w_{i-1}$ and so may be chosen for the ith basis vector w_i. An $(i-1)$-dimensional space is isomorphic to $\mathbb{F}^{i-1}$ and so has q^{i-1} elements. Thus there are $q^k - q^{i-1}$ choices for w_i. The process terminates when w_k is chosen, for then we have k linear independent vectors in a k-dimensional space, hence a basis.

(2) Let $\mathbb{F}$ be a finite field with q elements and let V be an n-dimensional vector space over $\mathbb{F}$. For each $k \in \{1, 2, \dots, n\}$ we show that the number of subspaces of V of dimension k is

$$\frac{(q^n - 1)(q^n - q)\dots(q^n - q^{k-1})}{(q^k - 1)(q^k - q)\dots(q^k - q^{k-1})}.$$

Any k-dimensional space is spanned by k independent vectors. By arguing as in the preceding example the numerator of the above expression is the number of ways of picking k independent vectors from an n-dimensional space. Two sets of k independent vectors span the same space W if and only if they are both bases of the k-dimensional space W. In order to obtain the formula for the number of distinct subspaces of dimension k we must divide by the number of repetitions, i.e. the number of bases of a fixed k-dimensional space. This factor which appears in the denominator is precisely the number computed in Example 1.

Next, we prove an important relation between the dimension of a subspace, the dimension of its associated quotient space and the dimension of the whole space:

Theorem 7. Let V be a vector space over F and let W be a subspace of V. Then V/W is a vector space with $\dim V = \dim W + \dim V/W$ (where if one side is infinite, both are).

Proof: Suppose W has dimension m and V has dimension n over F and let $w_1, w_2, \dots, w_m$ be a basis for W. By Corollary 5, these linearly independent elements of V can be extended to a basis $w_1, w_2, \dots, w_m, v_{m+1}, \dots, v_n$ of V. The natural surjective projection map of V into V/W maps each w_i to 0. No linear combination of the v_i is mapped to 0, since this would imply this linear combination is an element of W, contrary to the choice of the v_i. Hence, the image V/W of this projection map is isomorphic to the subspace of V spanned by the v_i, hence $\dim V/W = n - m$, which is the theorem when the dimensions are finite. If either side is infinite it is an easy exercise to produce an infinite number of linearly independent vectors showing the other side is also infinite.

Corollary 8. Let $\varphi : V \to U$ be a linear transformation of vector spaces over F. Then $\ker \varphi$ is a subspace of V, $\varphi(V)$ is a subspace of U and $\dim V = \dim \ker \varphi + \dim \varphi(V)$.

Proof: This follows immediately from Theorem 7. Note that the proof of Theorem 7 is in fact the special case of Corollary 8 where U is the quotient V/W and φ is the natural projection homomorphism.

Corollary 9. Let $\varphi : V \to W$ be a linear transformation of vector spaces of the same finite dimension. Then the following are equivalent:

(1) φ is an isomorphism

(2) φ is injective, i.e. $\ker \varphi = 0$

(3) φ is surjective, i.e. $\varphi(V) = W$

(4) φ sends a basis of V to a basis of W.

Proof: The equivalence of these conditions follows from Corollary 8 by counting dimensions.

Definition. If $\varphi : V \to U$ is a linear transformation of vector spaces over F, $\ker \varphi$ is sometimes called the *null space* of φ and the dimension of $\ker \varphi$ is called the *nullity* of φ. The dimension of $\varphi(V)$ is called the *rank* of φ. If $\ker \varphi = 0$, the transformation is said to be *nonsingular*.

Example

Let F be a finite field with q elements and let V be an n-dimensional vector space over F. Recall that the *general linear group* $GL(V)$ is the group of all nonsingular linear transformations from V to V (the group operation being composition). We show that the order of this group is

$$|GL(V)| = (q^n - 1)(q^n - q)(q^n - q^2)\cdots(q^n - q^{n-1}).$$

To see this, fix a basis $v_1, \ldots, v_n$ of V. A linear transformation is nonsingular if and only if it sends this basis to another basis of V. Moreover, if $w_1 \ldots, w_n$ is any basis of V, by Theorem 6 in Section 10.3 there is a unique linear transformation which sends v_i to w_i, $1 \le i \le n$. Thus the number of nonsingular linear transformations from V to itself equals the number of distinct bases of V. This number, which was computed in Example 1 above (with $k = n$), is the order of $GL(V)$.

EXERCISES

1. Let $V = \mathbb{R}^n$ and let $(a_1, a_2, \ldots, a_n)$ be a fixed vector in V. Prove that the collection of elements $(x_1, x_2, \ldots, x_n)$ of V with $a_1x_1 + a_2x_2 + \ldots + a_nx_n = 0$ is a subspace of V. Determine the dimension of this subspace and find a basis.
2. Let V be the collection of polynomials with coefficients in $\mathbb{Q}$ in the variable x of degree at most 5. Prove that V is a vector space over $\mathbb{Q}$ of dimension 6, with $1, x, x^2, \ldots, x^5$ as basis. Prove that $1, 1+x, 1+x+x^2, \ldots, 1+x+x^2+x^3+x^4+x^5$ is also a basis for V.
3. Let φ be the linear transformation $\varphi : \mathbb{R}^4 \to \mathbb{R}^1$ such that

$$\begin{array}{ll} \varphi((1,0,0,0)) = 1 & \varphi((1,-1,0,0)) = 0 \\ \varphi((1,-1,1,0)) = 1 & \varphi((1,-1,1,-1)) = 0. \end{array}$$

Determine $\varphi((a,b,c,d))$.

4. Prove that the space of real-valued functions on the closed interval $[a, b]$ is an infinite dimensional vector space.
5. Prove that the space of continuous real-valued functions on the closed interval $[a, b]$ is an infinite dimensional vector space.
6. Let V be a vector space of finite dimension. If φ is any linear transformation from V to V prove there is an integer m such that the intersection of the image of φ^m and the kernel of φ^m is $\{0\}$.
7. Let φ be a linear transformation from a vector space V of dimension n to itself that satisfies $\varphi^2 = 0$. Prove that the image of φ is contained in the kernel of φ and hence that the rank of φ is at most $n/2$.
8. Let V be a vector space over F and let φ be a linear transformation of the vector space V to itself. A nonzero element $v \in V$ satisfying $\varphi(v) = \lambda v$ for some $\lambda \in F$ is called an *eigenvector* of φ with *eigenvalue* λ. Prove that for any fixed $\lambda \in F$ the collection of eigenvectors of φ with eigenvalue λ together with 0 forms a subspace of V.
9. Let V be a vector space over F and let φ be a linear transformation of the vector space V to itself. Suppose for $i = 1, 2, \ldots, k$ that $v_i \in V$ is an eigenvector for φ with eigenvalue $\lambda_i \in F$ (cf. the preceding exercise) and that all the eigenvalues λ_i are distinct. Prove that $v_1, v_2, \ldots, v_k$ are linearly independent. [Use induction on k: write a linear dependence relation among the v_i and apply φ to get another linear dependence relation among the v_i involving the eigenvalues—now subtract a suitable multiple of the first linear relation to get a linear dependence relation on fewer elements.] Conclude that any linear transformation on an n-dimensional vector space has at most n distinct eigenvalues.

In the following exercises let V be a vector space of arbitrary dimension over a field F.

10. Prove that any vector space V has a basis (by convention the null set is the basis for the zero space). [Let $\mathcal{S}$ be the set of subsets of V consisting of linearly independent vectors, partially ordered under inclusion; apply Zorn's Lemma to $\mathcal{S}$ and show a maximal element of $\mathcal{S}$ is a basis].
11. Refine your argument in the preceding exercise to prove that any set of linearly independent vectors of V is contained in a basis of V.
12. If F is a field with a finite or countable number of elements and V is an infinite dimensional vector space over F with basis $\mathcal{B}$, prove that the cardinality of V equals the cardinality of $\mathcal{B}$. Deduce in this case that any two bases of V have the same cardinality.
13. Prove that as vector spaces over $\mathbb{Q}$, $\mathbb{R}^n \cong \mathbb{R}$, for all $n \in \mathbb{Z}^+$ (note that, in particular, this means $\mathbb{R}^n$ and $\mathbb{R}$ are isomorphic as additive abelian groups).
14. Let $\mathcal{A}$ be a basis for the infinite dimensional space V. Prove that V is isomorphic to the *direct sum* of copies of the field F indexed by the set $\mathcal{A}$. Prove that the *direct product* of copies of F indexed by $\mathcal{A}$ is a vector space over F and it has strictly larger dimension than the dimension of V (see the exercises in Section 10.3 for the definitions of direct sum and direct product of infinitely many modules).

11.2 THE MATRIX OF A LINEAR TRANSFORMATION

Throughout this section let V, W be vector spaces over the same field F, let $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ be an (ordered) basis of V, let $\mathcal{E} = \{w_1, w_2, \ldots, w_m\}$ be an (ordered) basis of W and let $\varphi \in \mathrm{Hom}(V, W)$ be a linear transformation from V to W. For each

$j \in \{1, 2, \ldots, n\}$ write the image of v_j under φ in terms of the basis $\mathcal{E}$:

$$\varphi(v_j) = \sum_{i=1}^{m} \alpha_{ij} w_i. \tag{11.3}$$

Let $M_{\mathcal{B}}^{\mathcal{E}}(\varphi) = (a_{ij})$ be the $m \times n$ matrix whose i, j entry is α_{ij} (that is, use the coefficients of the w_i's in the above computation of $\varphi(v_j)$ for the jth column of this matrix). The matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is called the *matrix of φ with respect to the bases $\mathcal{B}$, $\mathcal{E}$*. The domain basis is the lower and the codomain basis the upper letters appearing after the "M". Given this matrix, we can recover the linear transformation φ as follows: to compute $\varphi(v)$ for $v \in V$, write v in terms of the basis $\mathcal{B}$:

$$v = \sum_{i=1}^{n} \alpha_i v_i, \qquad \alpha_i \in F,$$

and then calculate the product of the $m \times n$ and $n \times 1$ matrices

$$M_{\mathcal{B}}^{\mathcal{E}}(\varphi) \times \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}.$$

The image of v under φ is given by

$$\varphi(v) = \sum_{i=1}^{m} \beta_i w_i \ ,$$

i.e. the column vector of coordinates of $\varphi(v)$ with respect to the basis $\mathcal{E}$ are obtained by multiplying the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ by the column vector of coordinates of v with respect to the basis $\mathcal{B}$ (sometimes denoted $[\varphi(v)]_{\mathcal{E}} = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)[v]_{\mathcal{B}}$).

Definition. The $m \times n$ matrix $A = (a_{ij})$ associated to the linear transformation φ above is said to *represent* the linear transformation φ with respect to the bases $\mathcal{B}$, $\mathcal{E}$. Similarly, φ is the linear transformation represented by A with respect to the bases $\mathcal{B}, \mathcal{E}$.

Examples

(1) Let $V = \mathbb{R}^3$ with the standard basis $\mathcal{B} = \{(1,0,0), (0,1,0), (0,0,1)\}$ and let $W = \mathbb{R}^2$ with the standard basis $\mathcal{E} = \{(1,0), (0,1)\}$. Let φ be the linear transformation $\varphi(x, y, z) = (x + 2y, x + y + z)$. Since $\varphi(1,0,0) = (1,1)$, $\varphi(0,1,0) = (2,1)$, $\varphi(0,0,1) = (0,1)$, the matrix $A = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is the matrix $\begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix}$.

(2) Let $V = W$ be the 2-dimensional space of solutions of the differential equation $y'' - 3y' + 2y = 0$ over $\mathbb{C}$ and let $\mathcal{B} = \mathcal{E}$ be the basis $v_1 = e^t$, $v_2 = e^{2t}$. Since the coefficients of this equation are constants it is easy to check that if y is a solution then its derivative y' is also a solution. It follows that the map $\varphi = d/dt =$ differentiation (with respect to t) is a linear transformation from V to itself. Since

$\varphi(v_1) = d(e^t)/dt = e^t = v_1$ and $\varphi(v_2) = d(e^{2t})/dt = 2e^{2t} = 2v_2$ we see that the corresponding matrix with respect to these bases is the diagonal matrix $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

(3) Let $V = W = \mathbb{Q}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{Q}\}$ be the usual 3-dimensional vector space of ordered 3-tuples with entries from the field $F = \mathbb{Q}$ of rational numbers and suppose φ is the linear transformation

$$\varphi(x, y, z) = (9x + 4y + 5z, -4x - 3z, -6x - 4y - 2z), \qquad x, y, z \in \mathbb{Q}$$

from V to itself. Take the standard basis $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$ for V and for $W = V$. Since $\varphi(1, 0, 0) = (9, -4, -6)$, $\varphi(0, 1, 0) = (4, 0, -4)$, $\varphi(0, 0, 1) = (5, -3, -2)$, the matrix A representing this linear transformation with respect to these bases is

$$A = \begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}.$$

Theorem 10. Let V be a vector space over F of dimension n and let W be a vector space over F of dimension m, with bases $\mathcal{B}$, $\mathcal{E}$ respectively. Then the map $\mathrm{Hom}_F(V, W) \to M_{m\times n}(F)$ from the space of linear transformations from V to W to the space of $m \times n$ matrices with coefficients in F defined by $\varphi \mapsto M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is a vector space isomorphism. In particular, there is a bijective correspondence between linear transformations and their associated matrices with respect to a fixed choice of bases.

Proof: The columns of the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ are determined by the action of φ on the basis $\mathcal{B}$ as in equation (3). This shows in particular that the map $\varphi \mapsto M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is an F-linear map since φ is F-linear. This map is *surjective* since given a matrix M, the map φ *defined* by equation (3) on a basis and then extended by linearity is a linear transformation with matrix M. The map is *injective* since two linear transformations agreeing on a basis are the same.

Note that different choices of bases give rise to different isomorphisms, so in the same sense that there is no natural choice of basis for a vector space, there is no natural isomorphism between $\mathrm{Hom}_F(V, W)$ and $M_{m\times n}(F)$.

Corollary 11. The dimension of $\mathrm{Hom}_F(V, W)$ is $(\dim V)(\dim W)$.

Proof: The dimension of $M_{m\times n}(F)$ is mn.

Definition. An $m \times n$ matrix A is called *nonsingular* if $Ax = 0$ with $x \in F^n$ implies $x = 0$.

The connection of the term *nonsingular* applied to matrices and to linear transformations is the following: let $A = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ be the matrix associated to the linear transformation φ (with some choice of bases $\mathcal{B}$, $\mathcal{E}$). Then independently of the choice of bases, the $m \times n$ matrix A is nonsingular if and only if the linear transformation φ is a nonsingular linear transformation from the n-dimensional space V to the m-dimensional space W (cf. the exercises).

Assume now that U, V and W are all finite dimensional vector spaces over F with ordered bases $\mathcal{D}$, $\mathcal{B}$ and $\mathcal{E}$ respectively, where $\mathcal{B}$ and $\mathcal{E}$ are as above and $\mathcal{D} = \{u_1, u_2, \ldots, u_k\}$. Assume $\psi : U \to V$ and $\varphi : V \to W$ are linear transformations. Their composite, $\varphi \circ \psi$, is a linear transformation from U to W, so we can compute its matrix with respect to the appropriate bases; namely, $M_{\mathcal{D}}^{\mathcal{E}}(\varphi \circ \psi)$ is found by computing

$$\varphi \circ \psi(u_j) = \sum_{i=1}^{m} \gamma_{ij} w_i$$

and putting the coefficients γ_{ij} down the jth column of $M_{\mathcal{D}}^{\mathcal{E}}(\varphi \circ \psi)$. Next, compute the matrices of ψ and φ separately:

$$\psi(u_j) = \sum_{p=1}^{n} \alpha_{pj} v_p \qquad \text{and} \qquad \varphi(v_p) = \sum_{i=1}^{m} \beta_{ip} w_i$$

so that $M_{\mathcal{D}}^{\mathcal{B}}(\psi) = (\alpha_{pj})$ and $M_{\mathcal{B}}^{\mathcal{E}}(\varphi) = (\beta_{ip})$.

Using these coefficients we can find an expression for the γ's in terms of the α's and β's as follows:

$$\begin{aligned}
\varphi \circ \psi(u_j) &= \varphi\left(\sum_{p=1}^{n} \alpha_{pj} v_p\right) \\
&= \sum_{p=1}^{n} \alpha_{pj} \varphi(v_p) \\
&= \sum_{p=1}^{n} \alpha_{pj} \sum_{i=1}^{m} \beta_{ip} w_i \\
&= \sum_{p=1}^{n} \sum_{i=1}^{m} \alpha_{pj} \beta_{ip} w_i.
\end{aligned}$$

By interchanging the order of summation in the above double sum we see that γ_{ij}, which is the coefficient of w_i in the above expression, is

$$\gamma_{ij} = \sum_{p=1}^{n} \alpha_{pj} \beta_{ip}.$$

Computing the product of the matrices for φ and ψ (in that order) we obtain

$$(\beta_{ij})(\alpha_{ij}) = (\delta_{ij}), \quad \text{where} \quad \delta_{ij} = \sum_{p=1}^{m} \beta_{ip} \alpha_{pj}.$$

By comparing the two sums above and using the commutativity of field multiplication, we see that for all i and j, $\gamma_{ij} = \delta_{ij}$. This computation proves the following result:

Theorem 12. With notations as above, $M_{\mathcal{D}}^{\mathcal{E}}(\varphi \circ \psi) = M_{\mathcal{B}}^{\mathcal{E}}(\varphi) M_{\mathcal{D}}^{\mathcal{B}}(\psi)$, i.e. with respect to a compatible choice of bases, the product of the matrices representing the

linear transformations φ and ψ is the matrix representing the composite linear transformation $\varphi \circ \psi$.

Corollary 13. Matrix multiplication is associative and distributive (whenever the dimensions are such as to make products defined). An $n \times n$ matrix A is nonsingular if and only if it is invertible.

Proof: Let A, B and C be matrices such that the products $(AB)C$ and $A(BC)$ are defined, and let S, T and R denote the associated linear transformations. By Theorem 12, the linear transformation corresponding to AB is the composite $S \circ T$ so the linear transformation corresponding to $(AB)C$ is the composite $(S \circ T) \circ R$. Similarly, the linear transformation corresponding to $A(BC)$ is the composite $S \circ (T \circ R)$. Since function composition is associative, these two linear transformations are the same, and so $(AB)C = A(BC)$ by Theorem 10. The distributivity is proved similarly. Note also that it is possible to prove these results by straightforward (albeit tedious) calculations with matrices.

If A is invertible, then $Ax = 0$ implies $x = A^{-1}Ax = A^{-1}0 = 0$, so A is nonsingular. Conversely, if A is nonsingular, fix bases $\mathcal{B}$, $\mathcal{E}$ for V and let φ be the linear transformation of V to itself represented by A with respect to these bases. By Corollary 9, φ is an isomorphism of V to itself, hence has an inverse, φ^{-1}. Let B be the matrix representing φ^{-1} with respect to the bases $\mathcal{E}$, $\mathcal{B}$ (note the order). Then $AB = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)M_{\mathcal{E}}^{\mathcal{B}}(\varphi^{-1}) = M_{\mathcal{E}}^{\mathcal{E}}(\varphi \circ \varphi^{-1}) = M_{\mathcal{E}}^{\mathcal{E}}(1) = I$, and B is the inverse of A.

Corollary 14

(1) If $\mathcal{B}$ is a basis of the n-dimensional space V, the map $\varphi \mapsto M_{\mathcal{B}}^{\mathcal{B}}(\varphi)$ is a ring and a vector space isomorphism of $\mathrm{Hom}_F(V, V)$ onto the space $M_n(F)$ of $n \times n$ matrices with coefficients in F.

(2) $GL(V) \cong GL_n(F)$ where $\dim V = n$. In particular, if F is a finite field the order of the finite group $GL_n(F)$ (which equals $|GL(V)|$) is given by the formula at the end of Section 1.

Proof: (1) We have already seen in Theorem 10 that this map is an isomorphism of vector spaces over F. Corollary 13 shows that $M_n(F)$ is a ring under matrix multiplication, and then Theorem 12 shows that multiplication is preserved under this map, hence it is also a ring isomorphism.

(2) This is immediate from (1) since a ring isomorphism sends units to units.

Definition. If A is any $m \times n$ matrix with entries from F, the *row rank* (respectively, *column rank*) of A is the number of linearly independent rows (respectively, columns) of A (where the rows or columns of A are considered as vectors in affine n-space, m-space, respectively).

The relation between the rank of a matrix and the rank of the associated linear transformation is the following: the rank of φ as a linear transformation equals the column rank of the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ (cf. the exercises). We shall also see that the row rank and the column rank of any matrix are the same.

We conclude this section by considering the relation of two matrices associated to the same linear transformation of a vector space to itself but with respect to two different choices of bases (cf. the exercises for the general statement regarding a linear transformation from a vector space V to another vector space W).

Definition. Two $n \times n$ matrices A and B are said to be *similar* if there is an invertible (i.e. nonsingular) $n \times n$ matrix P such that $P^{-1}AP = B$. Two linear transformations φ and ψ from a vector space V to itself are said to be *similar* if there is a nonsingular linear transformation ξ from V to V such that $\xi^{-1}\varphi\xi = \psi$.

Suppose $\mathcal{B}$ and $\mathcal{E}$ are two bases of the same vector space V and let $\varphi \in \mathrm{Hom}_F(V, V)$. Let I be the identity map from V to V and let $P = M_{\mathcal{E}}^{\mathcal{B}}(I)$ be its associated matrix (in other words, write the elements of the basis $\mathcal{E}$ in terms of the basis $\mathcal{B}$ — note the order — and use the resulting coordinates for the columns of the matrix P). Note that if $\mathcal{B} \neq \mathcal{E}$ then P is *not* the identity matrix. Then $P^{-1}M_{\mathcal{B}}^{\mathcal{B}}(\varphi)P = M_{\mathcal{E}}^{\mathcal{E}}(\varphi)$. If $[v]_{\mathcal{B}}$ is the $n \times 1$ matrix of coordinates for $v \in V$ with respect to the basis $\mathcal{B}$, and similarly $[v]_{\mathcal{E}}$ is the $n \times 1$ matrix of coordinates for $v \in V$ with respect to the basis $\mathcal{E}$, then $[v]_{\mathcal{B}} = P[v]_{\mathcal{E}}$. The matrix P is called the *transition* or *change of basis* matrix from $\mathcal{B}$ to $\mathcal{E}$ and this similarity action on $M_{\mathcal{B}}^{\mathcal{B}}(\varphi)$ is called a *change of basis*. This shows that the matrices associated to the same linear transformation with respect to two different bases are similar.

Conversely, suppose A and B are $n \times n$ matrices similar by a nonsingular matrix P. Let $\mathcal{B}$ be a basis for the n-dimensional vector space V. Define the linear transformation φ of V (with basis $\mathcal{B}$) to V (again with basis $\mathcal{B}$) by equation (3) using the given matrix A, i.e.

$$\varphi(v_j) = \sum_{i=1}^{n} \alpha_{ij} v_i.$$

Then $A = M_{\mathcal{B}}^{\mathcal{B}}(\varphi)$ by definition of φ. Define a new basis $\mathcal{E}$ of V by using the ith column of P for the coordinates of w_i in terms of the basis $\mathcal{B}$ (so $P = M_{\mathcal{E}}^{\mathcal{B}}(I)$ by definition). Then $B = P^{-1}AP = P^{-1}M_{\mathcal{B}}^{\mathcal{B}}(\varphi)P = M_{\mathcal{E}}^{\mathcal{E}}(\varphi)$ is the matrix associated to φ with respect to the basis $\mathcal{E}$. This shows that any two similar $n \times n$ matrices arise in this fashion as the matrices representing the same linear transformation with respect to two different choices of bases.

Note that change of basis for a linear transformation from V to itself is the same as conjugation by some element of the group $GL(V)$ of nonsingular linear transformations of V to V. In particular, the relation "similarity" is an equivalence relation whose equivalence classes are the orbits of $GL(V)$ acting by conjugation on $\mathrm{Hom}_F(V, V)$. If $\varphi \in GL(V)$ (i.e. φ is an invertible linear transformation), then the similarity class of φ is none other than the conjugacy class of φ in the group $GL(V)$.

Example

Let $V = \mathbb{Q}^3$ and let φ be the linear transformation

$$\varphi(x, y, z) = (9x + 4y + 5z, -4x - 3z, -6x - 4y - 2z), \qquad x, y, z \in \mathbb{Q}$$

from V to itself we considered in an earlier example. With respect to the standard basis, $\mathcal{B}$, $b_1 = (1,0,0)$, $b_2 = (0,1,0)$, $b_3 = (0,0,1)$ we saw that the matrix A representing this linear transformation is

$$A = M_{\mathcal{B}}^{\mathcal{B}}(\varphi) = \begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}.$$

Take now the basis, $\mathcal{E}$, $e_1 = (2,-1,-2)$, $e_2 = (1,0,-1)$, $e_3 = (3,-2,-2)$ for V (we shall see that this is in fact a basis momentarily). Since

$$\begin{aligned} \varphi(e_1) &= \varphi(2,-1,-2) = (4,-2,-4) = 2 \cdot e_1 + 0 \cdot e_2 + 0 \cdot e_3 \\ \varphi(e_2) &= \varphi(1,0,-1) = (4,-1,-4) = 1 \cdot e_1 + 2 \cdot e_2 + 0 \cdot e_3 \\ \varphi(e_3) &= \varphi(3,-2,-2) = (9,-6,-6) = 0 \cdot e_1 + 0 \cdot e_2 + 3 \cdot e_3, \end{aligned}$$

the matrix representing φ with respect to this basis is the matrix

$$B = M_{\mathcal{E}}^{\mathcal{E}}(\varphi) = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Writing the elements of the basis $\mathcal{E}$ in terms of the basis $\mathcal{B}$ we have

$$\begin{aligned} e_1 &= 2b_1 - b_2 - 2b_3 \\ e_2 &= b_1 - b_3 \\ e_3 &= 3b_1 - 2b_2 - 2b_3 \end{aligned}$$

so the matrix $P = M_{\mathcal{E}}^{\mathcal{B}}(I) = \begin{pmatrix} 2 & 1 & 3 \\ -1 & 0 & -2 \\ -2 & -1 & -2 \end{pmatrix}$ with inverse $P^{-1} = \begin{pmatrix} -2 & -1 & -2 \\ 2 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ conjugates A into B, i.e. $P^{-1}AP = B$, as can easily be checked. (Note incidentally that since P is invertible this proves that $\mathcal{E}$ is indeed a basis for V.)

We observe in passing that the matrix B representing this linear transformation φ is much simpler than the matrix A representing φ. The study of the simplest possible matrix representing a given linear transformation (and which basis to choose to realize it) is the study of *canonical forms* considered in the next chapter.

EXERCISES

1. Let V be the collection of polynomials with coefficients in $\mathbb{Q}$ in the variable x of degree at most 5. Determine the transition matrix from the basis $1, x, x^2, \ldots, x^5$ for V to the basis $1, 1+x, 1+x+x^2, \ldots, 1+x+x^2+x^3+x^4+x^5$ for V.
2. Let V be the vector space of the preceding exercise. Let $\varphi = d/dx$ be the linear transformation of V to itself given by usual differentiation of a polynomial with respect to x. Determine the matrix of φ with respect to the two bases for V in the previous exercise.
3. Let V be the collection of polynomials with coefficients in F in the variable x of degree at most n. Determine the transition matrix from the basis $1, x, x^2, \ldots, x^n$ for V to the elements

$$1,\ x-\lambda, \ldots,\ (x-\lambda)^{n-1},\ (x-\lambda)^n$$

where λ is a fixed element of F. Conclude that these elements are a basis for V.

4. Let φ be the linear transformation of $\mathbb{R}^2$ to itself given by rotation counterclockwise around the origin through an angle θ. Show that the matrix of φ with respect to the standard basis for $\mathbb{R}^2$ is $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.

5. Show that the $m \times n$ matrix A is nonsingular if and only if the linear transformation φ is a nonsingular linear transformation from the n-dimensional space V to the m-dimensional space W, where $A = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$, regardless of the choice of bases $\mathcal{B}$ and $\mathcal{E}$.

6. Prove if $\varphi \in \mathrm{Hom}_F(F^n, F^m)$, and $\mathcal{B}, \mathcal{E}$ are the natural bases of F^n, F^m respectively, then the range of φ equals the span of the set of columns of $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$. Deduce that the rank of φ (as a linear transformation) equals the column rank of $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$.

7. Prove that any two similar matrices have the same row rank and the same column rank.

8. Let V be an n–dimensional vector space over F and let φ be a linear transformation of the vector space V to itself.
   (a) Prove that if V has a basis consisting of eigenvectors for φ (cf. Exercise 8 of Section 1) then the matrix representing φ with respect to this basis (for both domain and range) is diagonal with the eigenvalues as diagonal entries.
   (b) If A is the $n \times n$ matrix representing φ with respect to a given basis for V (for both domain and range) prove that A is similar to a diagonal matrix if and only if V has a basis of eigenvectors for φ.

9. Suppose W is a subspace of the vector space V which is stable under the linear transformation φ (i.e. $\varphi(W) \leq W$). Show that φ induces a linear transformation $\bar{\varphi}$ on the quotient vector space V/W. Prove that φ is nonsingular on V if and only if φ restricted to W and $\bar{\varphi}$ on V/W are both nonsingular.

10. Let V be an n–dimensional vector space and let φ be a linear transformation of V to itself. Suppose W is a subspace of V of dimension m that is stable under φ.
    (a) Prove that there is a basis for V with respect to which the matrix for φ is of the form

$$\begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$

    where A is an $m \times m$ matrix, B is an $m \times (n-m)$ matrix and C is an $(n-m) \times (n-m)$ matrix (such a matrix is called *block upper triangular*).
    (b) Prove that if there is a subspace W' invariant under φ so that $V = W \oplus W'$ decomposes as a direct sum then the bases for W and W' give a basis for V with respect to which the matrix for φ is *block diagonal*:

$$\begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix}$$

    where A is an $m \times m$ matrix and C is an $(n-m) \times (n-m)$ matrix.
    (c) Prove conversely that if there is a basis for V with respect to which φ is block diagonal as in (b) then there are φ–invariant subspaces W and W' of dimensions m and $n-m$, respectively, with $V = W \oplus W'$.

11. Let φ be a linear transformation from the finite dimensional vector space V to itself such that $\varphi^2 = \varphi$.
    (a) Prove that image $\varphi \cap \ker \varphi = 0$.
    (b) Prove that $V = \text{image } \varphi \oplus \ker \varphi$.
    (c) Prove that there is a basis of V such that the matrix of φ with respect to this basis is a

diagonal matrix whose entries are all 0 or 1.

A linear transformation φ satisfying $\varphi^2 = \varphi$ is called an *idempotent* linear transformation. This exercise proves that idempotent linear transformations are simply projections onto some subspace.

12. Let $V = \mathbb{R}^2$, $v_1 = (1,0)$, $v_2 = (0,1)$, so that v_1, v_2 are a basis for V. Let φ be the linear transformation of V to itself whose matrix with respect to this basis is $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$. Prove that if W is the subspace generated by v_2 then W is stable under the action of φ. Prove that there is no subspace W' invariant under φ so that $V = W \oplus W'$.

13. Let V be a vector space of dimension n and let W be a vector space of dimension m over a field F. Suppose A is the $m \times n$ matrix representing a linear transformation φ from V to W with respect to the bases $\mathcal{B}_1$ for V and $\mathcal{E}_1$ for W. Suppose similarly that B is the $m \times n$ matrix representing φ with respect to the bases $\mathcal{B}_2$ for V and $\mathcal{E}_2$ for W. Let $P = M_{\mathcal{B}_2}^{\mathcal{B}_1}(I)$ where I denotes the identity map from V to V, and let $Q = M_{\mathcal{E}_2}^{\mathcal{E}_1}(I)$ where I denotes the identity map from W to W. Prove that $Q^{-1} = M_{\mathcal{E}_1}^{\mathcal{E}_2}(I)$ and that $Q^{-1}AP = B$, giving the general relation between matrices representing the same linear transformation but with respect to different choices of bases.

The following exercises recall the *Gauss–Jordan* elimination process. This is one of the fastest computational methods for the solution of a number of problems involving vector spaces — solving systems of linear equations, determining inverses of matrices, computing determinants, determining the span of a set of vectors, determining linear independence of a set of vectors etc.

Consider the system of m linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= c_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= c_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= c_m \end{aligned} \tag{11.4}$$

in the n unknowns $x_1, x_2, \ldots, x_n$ where $a_{ij}, c_i, i = 1, 2, \ldots, m, j = 1, 2, \ldots, n$ are elements of the field F. Associated to this system is the *coefficient matrix*:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

and the *augmented matrix*:

$$(A \mid C) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & c_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & c_m \end{array}\right)$$

(the term *augmented* refers to the presence of a column matrix $C = (c_i)$ in addition to the coefficient matrix $A = (a_{ij})$). The set of solutions in F of this system of equations is not

altered if we perform any of the following three operations:

(1) interchange any two equations
(2) add a multiple of one equation to another
(3) multiply any equation by a nonzero element from F,

which correspond to the following three *elementary row operations* on the augmented matrix:

(1) interchange any two rows
(2) add a multiple of one row to another
(3) multiply any row by a nonzero element from F.

If a matrix A can be transformed into the matrix C by a series of elementary row operations then A is said to be *row reduced* to C.

14. Prove that if A can be row reduced to C then C can be row reduced to A. Prove that the relation "$A \sim C$ if and only if A can be row reduced to C" is an equivalence relation. [Observe that the elementary row operations are reversible.]

Matrices lying in the same equivalence class under this equivalence relation are said to be *row equivalent*.

15. Prove that the row rank of two row equivalent matrices is the same. [It suffices to prove this for two matrices differing by an elementary row operation.]

An $m \times n$ matrix is said to be in *reduced row echelon form* if
(a) the first nonzero entry a_{ij_i} in row i is 1 and all other entries in the corresponding j_i^{th} column are zero, and
(b) $j_1 < j_2 < \ldots < j_r$ where r is the number of nonzero rows, i.e. the number of initial zeros in each row is strictly increasing (hence the term *echelon*).

An augmented matrix $(A \mid C)$ is said to be in reduced row echelon form if its coefficient matrix A is in reduced row echelon form. For example, the following two matrices are in reduced row echelon form:

$$\left(\begin{array}{cccccc|c} 1 & 0 & 5 & 7 & 0 & 3 & 0 \\ 0 & 1 & -1 & 1 & 0 & -4 & -1 \\ 0 & 0 & 0 & 0 & 1 & 6 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \quad \left(\begin{array}{cccc|c} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & -3 \end{array}\right)$$

(with $j_1 = 1$, $j_2 = 2$, $j_3 = 5$ for the first matrix and $j_1 = 2$, $j_2 = 4$ for the second matrix). The first nonzero entry in any given row of the coefficient matrix of a reduced row echelon augmented matrix (in position (i, j_i) by definition) is sometimes referred to as a *pivotal* element (so the pivotal elements in the first matrix are in positions (1,1), (2,2) and (3,5) and the pivotal elements in the second matrix are in positions (1,2) and (2,4)). The columns containing pivotal

elements will be called *pivotal* columns and the columns of the coefficient matrix not containing pivotal elements will be called *nonpivotal*.

16. Prove by induction that any augmented matrix can be put in reduced row echelon form by a series of elementary row operations.

17. Let A and C be two matrices in reduced row echelon form. Prove that if A and C are row equivalent then $A = C$.

18. Prove that the row rank of a matrix in reduced row echelon form is the number of nonzero rows.

19. Prove that the reduced row echelon forms of the matrices

$$\left(\begin{array}{cccccc|c} 1 & 1 & 4 & 8 & 0 & -1 & -1 \\ 1 & 2 & 3 & 9 & 0 & -5 & -2 \\ 0 & -2 & 2 & -2 & 1 & 14 & 3 \\ 1 & 4 & 1 & 11 & 0 & -13 & -4 \end{array}\right) \qquad \left(\begin{array}{cccc|c} 0 & -3 & 3 & 1 & 5 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 2 & -2 & 0 & -3 \end{array}\right)$$

are the two matrices preceding Exercise 16.

The point of the reduced row echelon form is that the corresponding system of linear equations is in a particularly simple form, from which the solutions to the system $AX = C$ in (4) can be determined immediately:

20. (*Solving Systems of Linear Equations*) Let $(A' \mid C')$ be the reduced row echelon form of the augmented matrix $(A \mid C)$. The number of zero rows of A' is clearly at least as great as the number of zero rows of $(A' \mid C')$.

(a) Prove that if the number of zero rows of A' is strictly larger than the number of zero rows of $(A' \mid C')$ then there are no solutions to $AX = C$.

By (a) we may assume that A' and $(A' \mid C')$ have the same number, r, of nonzero rows (so $n \geq r$).

(b) Prove that if $r = n$ then there is precisely one solution to the system of equations $AX = C$.

(c) Prove that if $r < n$ then there are infinitely many solutions to the system of equations $AX = C$. Prove in fact that the values of the $n - r$ variables corresponding to the nonpivotal columns of $(A' \mid C')$ can be chosen arbitrarily and that the remaining r variables corresponding to the pivotal columns of $(A' \mid C')$ are then determined uniquely.

21. Determine the solutions of the following systems of equations:

(a)

$$\begin{aligned} -3x + 3y + z &= 5 \\ x - y &= 0 \\ 2x - 2y &= -3 \end{aligned}$$

(b)

$$\begin{aligned} x - 2y + z &= 5 \\ x - 4y + 6z &= 10 \\ 4x - 11y + 11z &= 12 \end{aligned}$$

(c)

$$\begin{aligned} x - 2y + z &= 5 \\ y - 2z &= 17 \\ 2x - 3y &= 27 \end{aligned}$$

(d)

$$\begin{aligned} x + y - 3z + 2u &= 2 \\ 3x - 2y + 5z + u &= 1 \\ 6x + y - 4z + 3u &= 7 \\ 2x + 2y - 6z &= 4 \end{aligned}$$

(e)

$$\begin{aligned} x + y + 4z + 8u - w &= -1 \\ x + 2y + 3z + 9u - 5w &= -2 \\ -2y + 2z - 2u + v + 14w &= 3 \\ x + 4y + z + 11u - 13w &= -4. \end{aligned}$$

22. Suppose A and B are two row equivalent $m \times n$ matrices.

(a) Prove that the set

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

of solutions to the homogeneous linear equations $AX = 0$ as in equation (4) above are the same as the set of solutions to the homogeneous linear equations $BX = 0$. [It suffices to prove this for two matrices differing by an elementary row operation.]

(b) Prove that any linear dependence relation satisfied by the columns of A viewed as vectors in F^m is also satisfied by the columns of B.

(c) Conclude from (b) that the number of linearly independent columns of A is the same as the number of linearly independent columns of B.

23. Let A' be a matrix in reduced row echelon form.

(a) Prove that the nonzero rows of A' are linearly independent. Prove that the pivotal columns of A' are linearly independent and that the nonpivotal columns of A' are linearly dependent on the pivotal columns. (Note the role the pivotal elements play.)

(b) Prove that the number of linearly independent columns of a matrix in reduced row echelon form is the same as the number of linearly independent rows, i.e. the row rank and the column rank of such a matrix are the same.

24. Use the previous two exercises and Exercise 15 above to prove in general that the row rank and the column rank of a matrix are the same.

25. (*Computing Inverses of Matrices*) Let A be an $n \times n$ matrix.

(a) Show that A has an inverse matrix B with columns $B_1, B_2, \ldots, B_n$ if and only if the systems of equations:

$$AB_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad AB_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \ldots, \quad AB_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

have solutions.

(b) Prove that A has an inverse if and only if A is row equivalent to the $n \times n$ identity matrix.

(c) Prove that A has an inverse B if and only if the augmented matrix $(A \mid I)$ can be row reduced to the augmented matrix $(I \mid B)$ where I is the $n \times n$ identity matrix.

26. Determine the inverses of the following matrices using row reduction:

$$A = \begin{pmatrix} -7 & -1 & -4 \\ 7 & 1 & 3 \\ 1 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & -1 \\ 0 & 2 & 0 & 0 \\ -1 & 1 & 1 & 0 \end{pmatrix}.$$

27. (*Computing Spans, Linear Independence and Linear Dependencies in Vector Spaces*) Let V be an m–dimensional vector space with basis $e_1, e_2, \ldots, e_m$ and let $v_1, v_2, \ldots, v_n$ be vectors in V. Let A be the $m \times n$ matrix whose columns are the coordinates of the vectors v_i (with respect to the basis $e_1, e_2, \ldots, e_m$) and let A' be the reduced row echelon form of A.

(a) Let B be any matrix row equivalent to A. Let $w_1, w_2, \ldots, w_n$ be the vectors whose coordinates (with respect to the basis $e_1, e_2, \ldots, e_m$) are the columns of B. Prove that any linear relation

$$x_1 v_1 + x_2 v_2 + \ldots + x_n v_n = 0 \tag{11.5}$$

satisfied by $v_1, v_2, \ldots, v_n$ is also satisfied when v_i is replaced by w_i, $i = 1, 2, \ldots, n$.

(b) Prove that the vectors whose coordinates are given by the pivotal columns of A' are linearly independent and that the vectors whose coordinates are given by the nonpivotal columns of A' are linearly dependent on these.

(c) (*Determining Linear Independence of Vectors*) Prove that the vectors $v_1, v_2, \ldots, v_n$ are linearly independent if and only if A' has n nonzero rows (i.e. has rank n).

(d) (*Determining Linear Dependencies of Vectors*) By (c), the vectors $v_1, v_2, \ldots, v_n$ are linearly dependent if and only if A' has nonpivotal columns. The solutions to (5) defining linear dependence relations among $v_1, v_2, \ldots, v_n$ are given by the linear equations defined by A'. Show that each of the variables $x_1, x_2, \ldots, x_n$ in (5) corresponding to the nonpivotal columns of A' can be prescribed arbitrarily and the values of the remaining variables are then uniquely determined to give a linear dependence relation among $v_1, v_2, \ldots, v_n$ as in (5).

(e) (*Determining the Span of a set of vectors*) Prove that the subspace W spanned by $v_1, v_2, \ldots, v_n$ has dimension r where r is the number of nonzero rows of A' and that a basis for W is given by the original vectors v_{j_i} $(i = 1, 2, \ldots, r)$ corresponding to the pivotal columns of A'.

28. Let $V = \mathbb{R}^5$ with the standard basis and consider the vectors

$$v_1 = (1, 1, 3, -2, 3)\ ,\ v_2 = (0, 1, 0, -1, 0)\ ,\ v_3 = (2, 3, 6, -5, 6)$$
$$v_4 = (0, 3, 1, -3, 1)\ ,\ v_5 = (2, -1, -1, -1, -1)$$

(a) Show that the reduced row echelon form of the matrix

$$A = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 1 & 1 & 3 & 3 & -1 \\ 3 & 0 & 6 & 1 & -1 \\ -2 & -1 & -5 & -3 & -1 \\ 3 & 0 & 6 & 1 & -1 \end{pmatrix}$$

whose columns are the coordinates of v_1, v_2, v_3, v_4, v_5 is the matrix

$$A' = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 1 & 0 & 18 \\ 0 & 0 & 0 & 1 & -7 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the 1st, 2nd and 4th columns are pivotal and the remaining two are nonpivotal.

(b) Conclude that these vectors are linearly dependent, that the subspace W spanned by v_1, v_2, v_3, v_4, v_5 is 3–dimensional and that the vectors

$$v_1 = (1, 1, 3, -2, 3), \quad v_2 = (0, 1, 0, -1, 0) \quad \text{and} \quad v_4 = (0, 3, 1, -3, 1)$$

are a basis for W.

(c) Conclude from (a) that the coefficients x_1, x_2, x_3, x_4, x_5 of any linear relation

$$x_1 v_1 + x_2 v_2 + x_3 v_3 + x_4 v_4 + x_5 v_5 = 0$$

satisfied by v_1, v_2, v_3, v_4, v_5 are given by the equations

$$\begin{aligned} x_1 \quad + 2x_3 \quad + \ 2x_5 &= 0 \\ x_2 + \ x_3 \quad + 18x_5 &= 0 \\ x_4 - \ 7x_5 &= 0\,. \end{aligned}$$

Deduce that the 3rd and 5th variables, namely x_3 and x_5, corresponding to the nonpivotal columns of A', can be prescribed arbitrarily and the remaining variables are then uniquely determined as:

$$\begin{aligned} x_1 &= -2x_3 - 2x_5 \\ x_2 &= -x_3 - 18x_5 \\ x_4 &= 7x_5 \end{aligned}$$

to give all the linear dependence relations satisfied by v_1, v_2, v_3, v_4, v_5. In particular show that

$$-2v_1 - v_2 + v_3 = 0$$

and

$$-2v_1 - 18v_2 + 7v_4 + v_5 = 0$$

corresponding to $(x_3 = 1, x_5 = 0)$ and $(x_3 = 0, x_5 = 1)$, respectively.

29. For each exercise below, determine whether the given vectors in $\mathbb{R}^4$ are linearly independent. If they are linearly dependent, determine an explicit linear dependence among them.

(a) $(1, -4, 3, 0), (0, -1, 4, -3), (1, -1, 1, -1), (2, 2, -1, -3)$.

(b) $(1, -2, 4, 1), (2, -3, 9, -1), (1, 0, 6, -5), (2, -5, 7, 5)$.

(c) $(1, -2, 0, 1), (2, -2, 0, 0), (-1, 3, 0, -2), (-2, 1, 0, 1)$.

(d) $(0, 1, 1, 0), (1, 0, 1, 1), (2, 2, 2, 0), (0, -1, 1, 1)$.

30. For each exercise below, determine the subspace spanned in $\mathbb{R}^4$ by the given vectors and give a basis for this subspace.

(a) $(1, -2, 5, 3), (2, 3, 1, -4), (3, 8, -3, -5)$.

(b) $(2, -5, 3, 0), (0, -2, 5, -3), (1, -1, 1, -1), (-3, 2, -1, 2)$.

(c) $(1, -2, 0, 1), (2, -2, 0, 0), (-1, 3, 0, -2), (-2, 1, 0, 1)$.

(d) $(1, 1, 0, -1), (1, 2, 3, 0), (2, 3, 3, -1), (1, 2, 2, -2), (2, 3, 2, -3), (1, 3, 4, -3)$.

31. (*Computing the Image and Kernel of a Linear Transformation*) Let V be an n–dimensional vector space with basis $e_1, e_2, \ldots, e_n$ and let W be an m–dimensional vector space with basis $f_1, f_2, \ldots, f_m$. Let φ be a linear transformation from V to W and let A be the corresponding $m \times n$ matrix with respect to these bases: $A = (a_{ij})$ where

$$\varphi(e_j) = \sum_{i=1}^{m} a_{ij} f_i\,, \qquad j = 1, 2, \ldots, n,$$

i.e. the columns of A are the coordinates of the vectors $\varphi(e_1), \varphi(e_2), \ldots, \varphi(e_n)$ with respect to the basis $f_1, f_2, \ldots, f_m$ of W. Let A' be the reduced row echelon form of A.

(a) (*Determining the Image of a Linear Transformation*) Prove that the image $\varphi(V)$ of V under φ has dimension r where r is the number of nonzero rows of A' and that a basis for $\varphi(V)$ is given by the vectors $\varphi(e_{j_i})$ $(i = 1, 2, \ldots, r)$, i.e. the columns of A corresponding to the pivotal columns of A' give the coordinates of a basis for the image of φ.

(b) (*Determining the Kernel of a Linear Transformation*) The elements in the kernel of φ are the vectors in V whose coordinates $(x_1, x_2, \ldots, x_n)$ with respect to the basis $e_1, e_2, \ldots, e_n$ satisfy the equation

$$A \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = 0,$$

and the solutions $x_1, x_2, \ldots, x_n$ to this system of linear equations is determined by the matrix A'.

(i) Prove that φ is injective if and only if A' has n nonzero rows (i.e. has rank n).

(ii) By (i), the kernel of φ is nontrivial if and only if A' has nonpivotal columns. Show that each of the variables $x_1, x_2, \ldots, x_n$ above corresponding to the nonpivotal columns of A' can be prescribed arbitrarily and the values of the remaining variables are then uniquely determined to give an element $x_1e_1 + x_2e_2 + \ldots + x_ne_n$ in the kernel of φ. In particular, show that the coordinates of a basis for the kernel are obtained by successively setting one nonpivotal variable equal to 1 and all other nonpivotal variables to 0 and solving for the remaining pivotal variables. Conclude that the kernel of φ has dimension $n - r$ where r is the rank of A.

32. Let $V = \mathbb{R}^5$ and $W = \mathbb{R}^4$ with the standard bases. Let φ be the linear transformation $\varphi : V \to W$ defined by

$$\varphi(x, y, z, u, v) = (x + 2y + 3z + 4u + 4v, -2x - 4y + 2v, x + 2y + u - 2v, x + 2y - v).$$

(a) Prove that the matrix A corresponding to φ and these bases is

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 4 \\ -2 & -4 & 0 & 0 & 2 \\ 1 & 2 & 0 & 1 & -2 \\ 1 & 2 & 0 & 0 & -1 \end{pmatrix}$$

and that the reduced row echelon matrix A' row equivalent to A is

$$A' = \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the 1st, 3rd and 4th columns are pivotal and the remaining two are nonpivotal.

(b) Conclude that the image of φ is 3–dimensional and that the image of the 1st, 3rd and 4th basis elements of V, namely, $(1, -2, 1, 1)$, $(3, 0, 0, 0)$ and $(4, 0, 1, 0)$ give a basis for the image $\varphi(V)$ of V.

(c) Conclude from (a) that the elements in the kernel of φ are the vectors (x, y, z, u, v) satisfying the equations

$$\begin{aligned} x + 2y \qquad - v &= 0 \\ z + 3v &= 0 \\ u - v &= 0\,. \end{aligned}$$

Deduce that the 2nd and 5th variables, namely y and v, corresponding to the nonpivotal columns of A' can be prescribed arbitrarily and the remaining variables are then uniquely determined as:

$$\begin{aligned} x &= -2y + v \\ z &= -3v \\ u &= v\,. \end{aligned}$$

Show that $(-2, 1, 0, 0, 0)$ and $(1, 0, -3, 1, 1)$ give a basis for the 2–dimensional kernel of φ, corresponding to $(y = 1, v = 0)$ and $(y = 0, v = 1)$, respectively.

33. Let φ be the linear transformation from $\mathbb{R}^4$ to itself defined by the matrix

$$A = \begin{pmatrix} 1 & -1 & 0 & 3 \\ -1 & 2 & 1 & -1 \\ -1 & 1 & 0 & -3 \\ 1 & -2 & -1 & 1 \end{pmatrix}$$

with respect to the standard basis for $\mathbb{R}^4$. Determine a basis for the image and for the kernel of φ.

34. Let φ be the linear transformation $\varphi : \mathbb{R}^4 \to \mathbb{R}^2$ such that

$$\begin{aligned} &\varphi((1,0,0,0)) = (1,-1) \qquad && \varphi((1,-1,0,0)) = (0,0) \\ &\varphi((1,-1,1,0)) = (1,-1) && \varphi((1,-1,1,-1)) = (0,0). \end{aligned}$$

Determine a basis for the image and for the kernel of φ.

35. Let V be the set of all 2×2 matrices with real entries and let $\varphi : V \to \mathbb{R}$ be the map defined by sending a matrix $A \in V$ to the sum of the diagonal entries of A (the *trace* of A).

(a) Show that

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

is a basis for V.

(b) Prove that φ is a linear transformation and determine the matrix of φ with respect to the basis in (a) for V. Determine the dimension of and a basis for the kernel of φ.

36. Let V be the 6–dimensional vector space over $\mathbb{Q}$ consisting of the polynomials in the variable x of degree at most 5. Let φ be the map of V to itself defined by $\varphi(f) = x^2 f'' - 6xf' + 12f$, where f'' denotes the usual second derivative (with respect to x) of the polynomial $f \in V$ and f' similarly denotes the usual first derivative.

(a) Prove that φ is a linear transformation of V to itself.

(b) Determine a basis for the image and for the kernel of φ.

37. Let V be the 7–dimensional vector space over the field F consisting of the polynomials in the variable x of degree at most 6. Let φ be the linear transformation of V to itself defined by $\varphi(f) = f'$, where f' denotes the usual derivative (with respect to x) of the polynomial $f \in V$. For each of the fields below, determine a basis for the image and for the kernel

of φ:

(a) $F = \mathbb{R}$

(b) $F = \mathbb{F}_2$, the finite field of 2 elements (note that, for example, $(x^2)' = 2x = 0$ over this field)

(c) $F = \mathbb{F}_3$

(d) $F = \mathbb{F}_5$.

11.3 DUAL VECTOR SPACES

Definition

(1) For V any vector space over F let $V^* = \mathrm{Hom}_F(V, F)$ be the space of linear transformations from V to F, called the *dual space* of V. Elements of V^* are called *linear functionals*.

(2) If $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ is a basis of the finite dimensional space V, define $v_i^* \in V^*$ for each $i \in \{1, 2, \ldots, n\}$ by its action on the basis $\mathcal{B}$:

$$v_i^*(v_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \qquad 1 \leq j \leq n. \tag{11.6}$$

Proposition 15. With notations as above, $\{v_1^*, v_2^*, \ldots, v_n^*\}$ is a basis of V^*. In particular, if V is finite dimensional then V^* has the same dimension as V.

Proof: Observe that since V is finite dimensional, $\dim V^* = \dim \mathrm{Hom}_F(V, F) = \dim V = n$ (Corollary 11), so since there are n of the v_i^*'s it suffices to prove that they are linearly independent. If

$$\alpha_1 v_1^* + \alpha_2 v_2^* + \cdots + \alpha_n v_n^* = 0 \quad \text{in } \mathrm{Hom}_F(V, F),$$

then applying this element to v_i and using equation (6) above we obtain $\alpha_i = 0$. Since i is arbitrary these elements are linearly independent.

Definition. The basis $\{v_1^*, v_2^*, \ldots, v_n^*\}$ of V^* is called the *dual basis* to $\{v_1, v_2, \ldots, v_n\}$.

The exercises later show that if V is infinite dimensional it is always true that $\dim V < \dim V^*$. For spaces of arbitrary dimension the space V^* is the "algebraic" dual space to V. If V has some additional structure, for example a continuous structure (i.e. a topology), then one may define other types of dual spaces (*e.g.* the continuous dual of V, defined by requiring the linear functionals to be *continuous* maps). One has to be careful when reading other works (particularly analysis books) to ascertain what qualifiers are implicit in the use of the terms "dual space" and "linear functional".

Example

Let $[a, b]$ be a closed interval in $\mathbb{R}$ and let V be the real vector space of all continuous functions $f : [a, b] \to \mathbb{R}$. If $a < b$, V is infinite dimensional. For each $g \in V$ the function $\varphi_g : V \to \mathbb{R}$ defined by $\varphi_g(f) = \int_a^b f(t)g(t)dt$ is a linear functional on V.

Definition. The dual of $V^\star$, namely $V^{\star\star}$, is called the *double dual* or *second dual* of V.

Note that for a finite dimensional space V, $\dim V = \dim V^\star$ and also $\dim V^\star = \dim V^{\star\star}$, hence V and $V^{\star\star}$ are isomorphic vector spaces. For infinite dimensional spaces $\dim V < \dim V^{\star\star}$ (cf. the exercises) so V and $V^{\star\star}$ cannot be isomorphic. In the case of finite dimensional spaces there is a *natural*, i.e. basis independent or coordinate free way of exhibiting the isomorphism between a vector space and its second dual. The basic idea, in a more general setting, is as follows: if X is any set and S is any set of functions of X into our field F, we normally think of choosing or fixing an $f \in S$ and computing $f(x)$ as x ranges over all of X. Alternatively, we could think of fixing a point x in X and computing $f(x)$ as f ranges over all of S. The latter process, called *evaluation at* x shows that for each $x \in X$ there is a function $E_x : S \to F$ defined by $E_x(f) = f(x)$ (i.e. evaluate f at x). This gives a map $x \mapsto E_x$ of X into the set of F-valued functions on S. If S "separates points" in the sense that for distinct points x and y of X there is some $f \in S$ such that $f(x) \neq f(y)$, then the map $x \mapsto E_x$ is injective. The proof of the next lemma applies this "role reversal" process to the situation where $X = V$ and $S = V^\star$, proves E_x is a *linear* F-valued function on S, that is, E_x belongs to the dual space of $V^\star$, and proves the map $x \mapsto E_x$ is a linear transformation from V into $V^{\star\star}$. Note that throughout this process there is no mention of the word "basis" (although it is convenient to know the dimension of $V^{\star\star}$ — a fact we established by picking bases). In particular, the proof does not start with the familiar phrase "pick a basis of $V \ldots$".

Theorem 16. There is a natural injective linear transformation from V to $V^{\star\star}$. If V is finite dimensional then this linear transformation is an isomorphism.

Proof: Let $v \in V$. Define the map (*evaluation at* v)

$$E_v : V^\star \to F \qquad \text{by} \qquad E_v(f) = f(v).$$

Then $E_v(f + \alpha g) = (f + \alpha g)(v) = f(v) + \alpha g(v) = E_v(f) + \alpha E_g(v)$, so that E_v is a linear transformation from $V^\star$ to F. Hence E_v is an element of $\mathrm{Hom}_F(V^\star, F) = V^{\star\star}$. This defines a natural map

$$\varphi : V \to V^{\star\star} \qquad \text{by} \qquad \varphi(v) = E_v.$$

The map φ is a *linear* map, as follows: for $v, w \in V$ and $\alpha \in F$,

$$E_{v+\alpha w}(f) = f(v + \alpha x) = f(v) + \alpha f(w) = E_v(f) + \alpha E_w(f)$$

for every $f \in V^\star$, and so

$$\varphi(v + \alpha w) = E_{v+\alpha w} = E_v + \alpha E_w = \varphi(v) + \alpha\varphi(w).$$

To see that φ is injective let v be any nonzero vector in V. By the Building Up Lemma there is a basis $\mathcal{B}$ containing v. Let f be the linear transformation from V to F defined by sending v to 1 and every element of $\mathcal{B} - \{v\}$ to zero. Then $f \in V^\star$ and $E_v(f) = f(v) = 1$. Thus $\varphi(v) = E_v$ is not zero in $V^{\star\star}$. This proves $\ker \varphi = 0$, i.e. φ is injective.

If V has finite dimension n then by Proposition 15, V^* and hence also V^{**} has dimension n. In this case φ is an injective linear transformation from V to a finite dimensional vector space of the same dimension, hence is an isomorphism.

Let V, W be finite dimensional vector spaces over F with bases $\mathcal{B}, \mathcal{E}$, respectively and let $\mathcal{B}^*, \mathcal{E}^*$ be the dual bases. Fix some $\varphi \in \mathrm{Hom}_F(V, W)$. Then for each $f \in W^*$, the composite $f \circ \varphi$ is a linear transformation from V to F, that is $f \circ \varphi \in V^*$. Thus the map $f \mapsto f \circ \varphi$ defines a function from W^* to V^*. We denote this induced function on dual spaces by φ^*.

Theorem 17. With notations as above, φ^* is a linear transformation from W^* to V^* and $M_{\mathcal{E}^*}^{\mathcal{B}^*}(\varphi^*)$ is the transpose of the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ (recall that the transpose of the matrix (a_{ij}) is the matrix (a_{ji})).

Proof: The map φ^* is linear because $(f + \alpha g) \circ \varphi = (f \circ \varphi) + \alpha(g \circ \varphi)$. The equations which define φ are (from its matrix):

$$\varphi(v_j) = \sum_{i=1}^{m} \alpha_{ij} w_i \qquad 1 \le j \le n.$$

To compute the matrix for φ^*, observe that by the definitions of φ^* and w_k^*

$$\varphi^*(w_k^*)(v_j) = (w_k^* \circ \varphi)(v_j) = w_k^*\left(\sum_{i=1}^{m} \alpha_{ij} w_i\right) = \alpha_{kj}.$$

Also

$$(\sum_{i=1}^{n} \alpha_{ki} v_i^*)(v_j) = \alpha_{kj}$$

for all j. This shows that the the two linear functionals below agree on a basis of V, hence they are the same element of V^*:

$$\varphi^*(w_k^*) = \sum_{i=1}^{n} \alpha_{ki} v_i^*.$$

This determines the matrix for φ^* with respect to the bases $\mathcal{E}^*$ and $\mathcal{B}^*$ as the transpose of the matrix for φ.

Corollary 18. For any matrix A, the row rank of A equals the column rank of A.

Proof: Let $\varphi : V \to W$ be a linear transformation whose matrix with respect to some fixed bases of V and W is A. By Theorem 17 the matrix of $\varphi^* : W^* \to V^*$ with respect to the dual bases is the transpose of A. The column rank of A is the rank of φ and the row rank of A (= the column rank of the transpose of A) is the rank of φ^* (cf. Exercise 6 of Section 2). It therefore suffices to show that φ and φ^* have the same

rank. Now

$$f \in \ker \varphi^* \Leftrightarrow \varphi^*(f) = 0 \Leftrightarrow f \circ \varphi(v) = 0, \quad \text{for all } v \in V$$
$$\Leftrightarrow \varphi(V) \leq \ker f \Leftrightarrow f \in \mathrm{Ann}(\varphi(V)),$$

where $\mathrm{Ann}(V)$ is the annihilator of V described in Exercise 3 below. Thus $\mathrm{Ann}(\varphi(V)) = \ker \varphi^*$. By Exercise 3, $\dim \mathrm{Ann}(\varphi(V)) = \dim W - \dim \varphi(V)$. By Corollary 8, $\dim \ker \varphi^* = \dim W^* - \dim \varphi^*(W^*)$. Since W and W^* have the same dimension, $\dim \varphi(V) = \dim \varphi^*(W^*)$ as needed.

EXERCISES

1. Let V be a finite dimensional vector space. Prove that the map $\varphi \mapsto \varphi^*$ in Theorem 17 gives a ring isomorphism of $\mathrm{End}(V)$ with $\mathrm{End}(V^*)$.
2. Let V be the collection of polynomials with coefficients in $\mathbb{Q}$ in the variable x of degree at most 5 with $1, x, x^2, \ldots, x^5$ as basis. Prove that the following are elements of the dual space of V and express them as linear combinations of the dual basis:
   (a) $E : V \to \mathbb{Q}$ defined by $E(p(x)) = p(3)$ (i.e. evaluation at $x = 3$).
   (b) $\varphi : V \to \mathbb{Q}$ defined by $\varphi(p(x)) = \int_0^1 p(t)dt$.
   (c) $\varphi : V \to \mathbb{Q}$ defined by $\varphi(p(x)) = \int_0^1 t^2 p(t)dt$.
   (d) $\varphi : V \to \mathbb{Q}$ defined by $\varphi(p(x)) = p'(5)$ where $p'(x)$ denotes the usual derivative of the polynomial $p(x)$ with respect to x.
3. Let S be any subset of V^* for some finite dimensional space V. Define $\mathrm{Ann}(S) = \{v \in V \mid f(v) = 0 \text{ for all } f \in S\}$. ($\mathrm{Ann}(S)$ is called the *annihilator of S in V*).
   (a) Prove that $\mathrm{Ann}(S)$ is a subspace of V.
   (b) Let W_1 and W_2 be subspaces of V^*. Prove that $\mathrm{Ann}(W_1+W_2) = \mathrm{Ann}(W_1) \cap \mathrm{Ann}(W_2)$ and $\mathrm{Ann}(W_1 \cap W_2) = \mathrm{Ann}(W_1) + \mathrm{Ann}(W_2)$.
   (c) Let W_1 and W_2 be subspaces of V^*. Prove that $W_1 = W_2$ if and only if $\mathrm{Ann}(W_1) = \mathrm{Ann}(W_2)$.
   (d) Prove that the annihilator of S is the same as the annihilator of the subspace of V^* spanned by S.
   (e) Assume V is finite dimensional with basis $v_1, \ldots, v_n$. Prove that if $S = \{v_1^*, \ldots, v_k^*\}$ for some $k \leq n$, then $\mathrm{Ann}(S)$ is the subspace spanned by $\{v_{k+1}, \ldots, v_n\}$.
   (f) Assume V is finite dimensional. Prove that if W^* is any subspace of V^* then $\dim \mathrm{Ann}(W^*) = \dim V - \dim W^*$.
4. If V is infinite dimensional with basis $\mathcal{A}$, prove that $\mathcal{A}^* = \{v^* \mid v \in \mathcal{A}\}$ does *not* span V^*.
5. If V is infinite dimensional with basis $\mathcal{A}$, prove that V^* is isomorphic to the direct product of copies of F indexed by $\mathcal{A}$. Deduce that $\dim V^* > \dim V$. [Use Exercise 14 of Section 1]

11.4 DETERMINANTS

Although we shall be using the theory primarily for vector spaces over a field, the theory of determinants can be developed with no extra effort over arbitrary commutative rings with 1. Thus in this section R is any commutative ring with 1 and $V_1, V_2, \ldots, V_n$, V and W are R-modules.

Definition

(1) A map $\varphi : V_1 \times V_2 \times \cdots \times V_n \to W$ is called *multilinear* if for each fixed i and fixed elements $v_j \in V_j$, $j \neq i$, the map

$$V_i \to W \qquad \text{defined by} \qquad x \mapsto \varphi(v_1, \ldots, v_{i-1}, x, v_{i+1}, \ldots, v_n)$$

is an R-module homomorphism. If $V_i = V$, $i = 1, 2, \ldots, n$, then φ is called an *n-multilinear function on V*, and if in addition $W = R$, φ is called an *n-multilinear form on V*.

(2) An n-multilinear function φ on V is called *alternating* if $\varphi(v_1, v_2, \ldots, v_n) = 0$ whenever $v_i = v_{i+1}$ for some $i \in \{1, 2, \ldots, n-1\}$ (i.e. φ is zero whenever two consecutive arguments are equal). The function φ is called *symmetric* if interchanging v_i and v_j for any i and j in $(v_1, v_2, \ldots, v_n)$ does not alter the value of φ on this n-tuple.

When $n = 2$ (respectively, 3) one says φ is *bilinear* (respectively *trilinear*) rather than 2-multilinear (respectively 3-multilinear). Also, when n is clear from the context we shall simply say φ is multilinear.

Example

For any fixed $m \geq 0$ the usual dot product on $V = \mathbb{R}^m$ is a bilinear form (here the ring R is the field of real numbers).

Proposition 19. Let φ be an n-multilinear alternating function on V. Then

(1) $\varphi(v_1, \ldots, v_{i-1}, v_{i+1}, v_i, v_{i+2}, \ldots, v_n) = -\varphi(v_1, v_2, \ldots, v_n)$ for any $i \in \{1, 2, \ldots, n-1\}$, i.e. the value of φ on an n-tuple is negated if two adjacent components are interchanged.

(2) For each $\sigma \in S_n$, $\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \epsilon(\sigma)\varphi(v_1, v_2, \ldots, v_n)$, where $\epsilon(\sigma)$ is the sign of the permutation σ (cf. Section 3.5).

(3) If $v_i = v_j$ for any pair of distinct $i, j \in \{1, 2, \ldots, n-1\}$ then $\varphi(v_1, v_2, \ldots, v_n) = 0$.

(4) If v_i is replaced by $v_i + \alpha v_j$ in $(v_1, \ldots, v_n)$ for any $j \neq i$ and any $\alpha \in R$, the value of φ on this n-tuple is not changed.

Proof: (1) Let $\psi(x, y)$ be the function φ with variable entries x and y in positions i and $i + 1$ respectively and fixed entries v_j in position j, for all other j. Thus (1) is the same as showing $\psi(y, x) = -\psi(x, y)$. Since φ is alternating $\psi(x + y, x + y) = 0$. Expanding $x + y$ in each variable in turn gives $\psi(x + y, x + y) = \psi(x, x) + \psi(x, y) + \psi(y, x) + \psi(y, y)$. Again, by the alternating property of φ, the first and last terms on the right hand side of the latter equation are zero. Thus $0 = \psi(x, y) + \psi(y, x)$, which gives (1).

(2) Every permutation can be written as a product of transpositions (cf. Section 3.5). Furthermore, every transposition may be written as a product of transpositions which interchange two successive integers (cf. Exercise 6 of Section 3.5). Thus every permutation σ can be written as $\tau_1 \cdots \tau_m$, where τ_k is a transposition interchanging two successive integers, for all k. It follows from m applications of (1) that

$$\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \epsilon(\tau_m) \cdots \epsilon(\tau_1)\varphi(v_1, v_2, \ldots, v_n).$$

Finally, since ϵ is a homomorphism into the abelian group ± 1 (so the order of the factors ± 1 does not matter), $\epsilon(\tau_1)\cdots\epsilon(\tau_m) = \epsilon(\tau_1\cdots\tau_m) = \epsilon(\sigma)$. This proves (2).

(3) Choose σ to be any permutation which fixes i and moves j to $i+1$. Thus $(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)})$ has two equal adjacent components so φ is zero on this n-tuple. By (2), $\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \pm\varphi(v_1, v_2, \ldots, v_n)$. This implies (3).

(4) This follows immediately from (3) on expanding by linearity in the ith position.

Proposition 20. Assume φ is an n-multilinear alternating function on V and that for some $v_1, v_2, \ldots, v_n$ and $w_1, w_2, \ldots, w_n \in V$ and some $\alpha_{ij} \in R$ we have

$$\begin{aligned} w_1 &= \alpha_{11}v_1 + \alpha_{21}v_2 + \cdots + \alpha_{n1}v_n \\ w_2 &= \alpha_{12}v_1 + \alpha_{22}v_2 + \cdots + \alpha_{n2}v_n \\ &\vdots \\ w_n &= \alpha_{1n}v_1 + \alpha_{2n}v_2 + \cdots + \alpha_{nn}v_n \end{aligned}$$

(we have purposely written the indices of the α_{ij} in "column format"). Then

$$\varphi(w_1, w_2, \ldots, w_n) = \sum_{\sigma\in S_n} \epsilon(\sigma)\alpha_{\sigma(1)\,1}\alpha_{\sigma(2)\,2}\cdots\alpha_{\sigma(n)\,n}\varphi(v_1, v_2, \ldots, v_n).$$

Proof: If we expand $\varphi(w_1, w_2, \ldots, w_n)$ by multilinearity we obtain a sum of n^n terms of the form $\alpha_{i_1\,1}\alpha_{i_2\,2}\ldots\alpha_{i_n\,n}\varphi(v_{i_1}, v_{i_2}, \ldots, v_{i_n})$, where the indices $i_1, i_2, \ldots, i_n$ each run over $1, 2, \ldots, n$. By Proposition 19(3), φ is zero on the terms where two or more of the i_j's are equal. Thus in this expansion we need only consider the terms where $i_1, \ldots, i_n$ are distinct. Such sequences are in bijective correspondence with permutations in S_n, so each nonzero term may be written as $\alpha_{\sigma(1)\,1}\alpha_{\sigma(2)\,2}\cdots\alpha_{\sigma(n)\,n}\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)})$, for some $\sigma \in S_n$. Applying (2) of the previous proposition to each of these terms in the expansion of $\varphi(w_1, w_2, \ldots, w_n)$ gives the expression in the proposition.

Definition. An $n \times n$ *determinant function* on R is any function

$$\det : M_{n\times n}(R) \to R$$

that satisfies the following two axioms:

(1) det is an n-multilinear alternating form on $R^n (= V)$, where the n-tuples are the n columns of the matrices in R^n

(2) $\det(I) = 1$, where I is the $n \times n$ identity matrix.

On occasion we shall write $\det(A_1, A_2, \ldots, A_n)$ for $\det A$, where $A_1, A_2, \ldots, A_n$ are the columns of A.

Theorem 21. There is a unique $n \times n$ determinant function on R and it can be computed for any $n \times n$ matrix (α_{ij}) by the formula:

$$\det(\alpha_{ij}) = \sum_{\sigma\in S_n} \epsilon(\sigma)\alpha_{\sigma(1)\,1}\alpha_{\sigma(2)\,2}\cdots\alpha_{\sigma(n)\,n}.$$

Proof: Let $A_1, A_2, \ldots, A_n$ be the column vectors in a general $n \times n$ matrix (α_{ij}). We leave it as an exercise to check that the formula given in the statement of the theorem does satisfy the axioms of a determinant function — this gives existence of a determinant function. To prove uniqueness let e_i be the column n-tuple with 1 in position i and zeros in all other positions. Then

$$\begin{aligned} A_1 &= \alpha_{11}e_1 + \alpha_{21}e_2 + \cdots + \alpha_{n1}e_n \\ A_2 &= \alpha_{12}e_1 + \alpha_{22}e_2 + \cdots + \alpha_{n2}e_n \\ &\vdots \\ A_n &= \alpha_{1n}e_1 + \alpha_{2n}e_2 + \cdots + \alpha_{nn}e_n. \end{aligned}$$

By Proposition 20, $\det A = \sum_{\sigma \in S_n} \epsilon(\sigma)\alpha_{\sigma(1)\,1}\alpha_{\sigma(2)\,2} \cdots \alpha_{\sigma(n)\,n} \det(e_1, e_2, \ldots, e_n)$. Since by axiom (2) of a determinant function $\det(e_1, e_2, \ldots, e_n) = 1$, the value of $\det A$ is as claimed.

Corollary 22. The determinant is an n-multilinear function of the rows of $M_{n\times n}(R)$ and for any $n \times n$ matrix A, $\det A = \det(A^t)$, where A^t is the transpose of A.

Proof: The first statement is an immediate consequence of the second, so it suffices to prove that a matrix and its transpose have the same determinant. For $A = (\alpha_{ij})$ one calculates that

$$\det A^t = \sum_{\sigma \in S_n} \epsilon(\sigma)\alpha_{1\,\sigma(1)}\alpha_{2\,\sigma(2)} \cdots \alpha_{n\,\sigma(n)}.$$

Each number from 1 to n appears exactly once among $\sigma(1), \ldots, \sigma(n)$ so we may rearrange the product $\alpha_{1\,\sigma(1)}\alpha_{2\,\sigma(2)} \cdots \alpha_{n\,\sigma(n)}$ as $\alpha_{\sigma^{-1}(1)\,1}\alpha_{\sigma^{-1}(2)\,2} \cdots \alpha_{\sigma^{-1}(n)\,n}$. Also, the homomorphism ϵ takes values in $\{\pm 1\}$ so $\epsilon(\sigma) = \epsilon(\sigma^{-1})$. Thus the sum for $\det A^t$ may be rewritten as

$$\sum_{\sigma \in S_n} \epsilon(\sigma^{-1})\alpha_{\sigma^{-1}(1)\,1}\alpha_{\sigma^{-1}(2)\,2} \cdots \alpha_{\sigma^{-1}(n)\,n}.$$

The latter sum is over all permutations, so the index σ^{-1} may be replaced by σ. The resulting expression is the sum for $\det A$. This completes the proof.

Theorem 23. (Cramer's Rule) If $A_1, A_2, \ldots, A_n$ are the columns of an $n \times n$ matrix A and $B = \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_n A_n$, for some $\beta_1, \ldots, \beta_n \in R$, then

$$\beta_i \det A = \det(A_1, \ldots, A_{i-1}, B, A_{i+1}, \ldots, A_n).$$

Proof: This follows immediately from Proposition 19(3) on replacing the given expression for B in the ith position and expanding by multilinearity in that position.

Corollary 24. If R is an integral domain, then $\det A = 0$ for $A \in M_n(R)$ if and only if the columns of A are R-linearly dependent as elements of the free R-module of rank n. Also, $\det A = 0$ if and only if the rows of A are R-linearly dependent.

Proof: Since $\det A = \det A^t$ the first sentence implies the second. Assume first that the columns of A are linearly dependent and

$$0 = \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_n A_n$$

is a dependence relation on the columns of A with, say, $\beta_i \neq 0$. By Cramer's rule, $\beta_i \det A = 0$. Since R is an integral domain and $\beta_i \neq 0$, $\det A = 0$.

Conversely, assume the columns of A are independent. Consider the integral domain R as embedded in its quotient field F so that $M_{n\times n}(R)$ may be considered as a subring of $M_{n\times n}(F)$ (and note that the determinant function on the subring is the restriction of the determinant function from $M_{n\times n}(F)$). The columns of A in this way become elements of F^n. Any nonzero F-linear combination of the columns of A which is zero in F^n gives, by multiplying the coefficients by a common denominator, a nonzero R-linear dependence relation. The columns of A must therefore be independent vectors in F^n. Since A has n columns, these form a basis of F^n. Thus there are elements β_{ij} of F such that for each i, the ith basis vector e_i in F^n may be expressed as

$$e_i = \beta_{1i} A_1 + \beta_{2i} A_2 + \cdots + \beta_{ni} A_n.$$

The $n \times n$ identity matrix is the one whose columns are $e_1, e_2, \ldots, e_n$. By Proposition 20 (with $\varphi = \det$), the determinant of the identity matrix is some F-multiple of $\det A$. Since the determinant of the identity matrix is 1, $\det A$ cannot be zero. This completes the proof.

Theorem 25. For matrices $A, B \in M_{n\times n}(R)$, $\det AB = (\det A)(\det B)$.

Proof: Let $B = (\beta_{ij})$ and let $A_1, A_2, \ldots, A_n$ be the columns of A. Then $C = AB$ is the $n \times n$ matrix whose j^{th} column is $C_j = \beta_{1j} A_1 + \beta_{2j} A_2 + \cdots + \beta_{nj} A_n$. By Proposition 20 applied to the multilinear function det we obtain

$$\det C = \det(C_1, \ldots, C_n) = \left[\sum_{\sigma \in S_n} \epsilon(\sigma) \beta_{\sigma(1)\,1} \beta_{\sigma(2)\,2} \cdots \beta_{\sigma(n)\,n} \right] \det(A_1, \ldots, A_n).$$

The sum inside the brackets is the formula for $\det B$, hence $\det C = (\det B)(\det A)$, as required ($R$ is commutative).

Definition. Let $A = (\alpha_{ij})$ be an $n \times n$ matrix. For each i, j, let A_{ij} be the $n-1 \times n-1$ matrix obtained from A by deleting its ith row and j^{th} column (an $n-1 \times n-1$ *minor* of A). Then $(-1)^{i+j} \det(A_{ij})$ is called the ij *cofactor of* A.

Theorem 26. (The Cofactor Expansion Formula along the ith row) If $A = (\alpha_{ij})$ is an $n \times n$ matrix, then for each fixed $i \in \{1, 2, \ldots, n\}$ the determinant of A can be computed from the formula

$$\det A = (-1)^{i+1} \alpha_{i1} \det A_{i1} + (-1)^{i+2} \alpha_{i2} \det A_{i2} + \cdots + (-1)^{i+n} \alpha_{in} \det A_{in}.$$

Proof: For each A let $D(A)$ be the element of R obtained from the cofactor expansion formula described above. We prove that D satisfies the axioms of a determinant function, hence is *the* determinant function. Proceed by induction on n. If

$n = 1$, $D((\alpha)) = \alpha$, for all 1×1 matrices (α) and the result holds. Assume therefore that $n \geq 2$. To show that D is an alternating multilinear function of the columns, fix an index k and consider the kth column as varying and all other columns as fixed. If $j \neq k$, α_{ij} does not depend on k and $D(A_{ij})$ is linear in the kth column by induction. Also, as the kth column varies linearly so does α_{ik}, whereas $D(A_{ik})$ remains unchanged (the kth column has been deleted from A_{ik}). Thus each term in the formula for D varies linearly in the kth column. This proves D is multilinear in the columns.

To prove D is alternating assume columns k and $k+1$ of A are equal. If $j \neq k$ or $k+1$, the two equal columns of A become two equal columns in the matrix A_{ij}. By induction $D(A_{ij}) = 0$. The formula for D therefore has at most two nonzero terms: when $j = k$ and when $j = k+1$. The minor matrices A_{ik} and $A_{i\,k+1}$ are identical and $a_{ik} = a_{i\,k+1}$. Then the two remaining terms in the expansion for D, $(-1)^{i+k} a_{ik} D(A_{ik})$ and $(-1)^{i+k+1} a_{i\,k+1} D(A_{i\,k+1})$ are equal and appear with opposite signs, hence they cancel. Thus $D(A) = 0$ if A has two adjacent columns which are equal, i.e. D is alternating.

Finally, it follows easily from the formula and induction that $D(I) = 1$, where I is the identity matrix. This completes the induction.

Theorem 27. (Cofactor Formula for the Inverse of a Matrix) Let $A = (\alpha_{ij})$ be an $n \times n$ matrix and let B be the transpose of its matrix of cofactors, i.e. $B = (\beta_{ij})$ where $\beta_{ij} = (-1)^{i+j} \det A_{ji}$, $1 \leq i, j \leq n$. Then $AB = BA = (\det A)I$. Moreover, $\det A$ is a unit in R if and only if A is a unit in $M_{n\times n}(R)$; in this case the matrix $\dfrac{1}{\det A} B$ is the inverse of A.

Proof: The i, j entry of AB is $\alpha_{i1}\beta_{1j} + \alpha_{i2}\beta_{2j} + \cdots + \alpha_{in}\beta_{nj}$. By definition of the entries of B this equals

$$\alpha_{i1}(-1)^{j+1} D(A_{j1}) + \alpha_{i2}(-1)^{j+2} D(A_{j2}) + \cdots + \alpha_{in}(-1)^{j+n} D(A_{jn}). \tag{11.7}$$

If $i = j$, this is the cofactor expansion for $\det A$ along the ith row. The diagonal entries of AB are thus all equal to $\det A$. If $i \neq j$, let $\overline{A}$ be the matrix A with the jth row replaced by the ith row, so $\det \overline{A} = 0$. By inspection $\overline{A}_{jk} = A_{jk}$ and $\alpha_{ik} = \overline{\alpha}_{jk}$ for every $k \in \{1, 2, \ldots, n\}$. By making these substitutions in equation (7) for each $k = 1, 2, \ldots, n$ one sees that the i, j entry in AB equals $\overline{\alpha}_{j1}(-1)^{1+j} D(\overline{A}_{j1}) + \cdots + \overline{\alpha}_{jn}(-1)^{n+j} D(\overline{A}_{jn})$. This expression is the cofactor expansion for $\det \overline{A}$ along the jth row. Since, as noted above, $\det \overline{A} = 0$, this proves that all off diagonal terms of AB are zero, i.e. $AB = (\det A)I$.

It follows directly from the definition of B that the pair (A^t, B^t) satisfies the same hypotheses as the pair (A, B). The preceding result therefore shows that $(BA)^t = A^t B^t = (\det A^t)I$. Since $\det A^t = \det A$ and the transpose of a diagonal matrix is itself, we obtain $BA = (\det A)I$ as well.

If $d = \det A$ is a unit in R, then $d^{-1}B$ is a matrix with entries in R whose product with A (on either side) is the identity, i.e. A is a unit in $M_{n\times n}(R)$. Conversely, assume that A is a unit in R with (2-sided) inverse matrix B. Since $\det B \in R$ and $1 = \det I = \det AB = (\det A)(\det B) = (\det B)(\det A)$, $\det A$ has a 2-sided inverse in R, as needed. This completes all parts of the proof.

EXERCISES

1. Formulate and prove the cofactor expansion formula along the jth column of a square matrix A.
2. Let F be a field and let $A_1, A_2, \ldots, A_n$ be (column) vectors in F^n. Form the matrix A whose i^{th} column is A_i. Prove that these vectors form a basis of F^n if and only if $\det A \neq 0$.
3. Let R be any commutative ring with 1, let V be an R-module and let $x_1, x_2, \ldots, x_n \in V$. Assume that for some $A \in M_{n\times n}(R)$,

$$A\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = 0.$$

   Prove that $(\det A)x_i = 0$, for all $i \in \{1, 2, \ldots, n\}$.
4. (*Computing Determinants of Matrices*) This exercise outlines the use of Gauss–Jordan elimination (cf. the exercises in Section 2) to compute determinants. This is the most efficient general procedure for computing large determinants. Let A be an $n \times n$ matrix.
   (a) Prove that the elementary row operations have the following effect on determinants:
      (i) interchanging two rows changes the sign of the determinant
      (ii) adding a multiple of one row to another does not alter the determinant
      (iii) multiplying any row by a nonzero element u from F multiplies the determinant by u.
   (b) Prove that $\det(A)$ is nonzero if and only if A is row equivalent to the $n \times n$ identity matrix. Suppose A can be row reduced to the identity matrix using a total of s row interchanges as in (i) and by multiplying rows by the nonzero elements $u_1, u_2, \ldots, u_t$ as in (iii). Prove that $\det(A) = (-1)^s(u_1u_2\ldots u_t)^{-1}$.
5. Compute the determinants of the following matrices using row reduction:

$$A = \begin{pmatrix} 5 & 4 & -6 \\ -2 & 0 & 2 \\ 3 & 4 & -2 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix}.$$

12

Modules Over Principal Ideal Domains

The main purpose in this chapter is to prove a structure theorem for finitely generated modules over particularly nice rings, namely Principal Ideal Domains. This theorem is an example of the ideal structure of the ring (which is particularly simple for P.I.D.s) being reflected in the structure of its modules. If we apply this result in the case where the P.I.D. is the ring of integers $\mathbb{Z}$ then we obtain a proof of the Fundamental Theorem of Finitely Generated Abelian Groups (which we examined in Chapter 5 without proof). If instead we apply this structure theorem in the case where the P.I.D. is the ring $F[x]$ of polynomials in x with coefficients in a field F we shall obtain the basic results on the so-called rational and Jordan canonical forms for a matrix. Before proceeding to the proof we briefly discuss these two important applications.

We have already discussed in Chapter 5 the result that any finitely generated abelian group is isomorphic to the direct sum of cyclic abelian groups, either $\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$ for some positive integer $n \neq 0$. Recall also that an abelian group is the same thing as a $\mathbb{Z}$-module. Since the ideals of $\mathbb{Z}$ are precisely the trivial ideal (0) and the principal ideals $(n) = n\mathbb{Z}$ generated by positive integers n, we see that the Fundamental Theorem of Finitely Generated Abelian Groups in the language of modules says that any finitely generated $\mathbb{Z}$-module is the direct sum of modules of the form $\mathbb{Z}/I$ where I is an ideal of $\mathbb{Z}$ (these are the cyclic $\mathbb{Z}$-modules), together with a uniqueness statement when the direct sum is written in a particular form. Note the correspondence between the ideal structure of $\mathbb{Z}$ and the structure of its (finitely generated) modules, the finitely generated abelian groups.

The Fundamental Theorem of Finitely Generated Modules over a P.I.D. states that the same result holds when the Principal Ideal Domain $\mathbb{Z}$ is replaced by *any* P.I.D. In particular, we have seen in Chapter 10 that a module over the ring $F[x]$ of polynomials in x with coefficients in the field F is the same thing as a vector space V together with a fixed linear transformation T of V (where the element x acts on V by the linear transformation T). The Fundamental Theorem in this case will say that such a vector space is the direct sum of modules of the form $F[x]/I$ where I is an ideal of $F[x]$, hence is either the trivial ideal (0) or a principal ideal $(f(x))$ generated by some nonzero polynomial $f(x)$ (these are the cyclic $F[x]$-modules), again with a uniqueness statement when the direct sum is written in a particular form. If this is translated back into the language of vector spaces and linear transformations we can obtain information on the linear transformation T.

For example, suppose V is a vector space of dimension n over F and we choose a basis for V. Then giving a linear transformation T of V to itself is the same thing as giving an $n \times n$ matrix A with coefficients in F (and choosing a different basis for V gives a different matrix B for T which is similar to A i.e. is of the form $P^{-1}AP$ for some invertible matrix P which defines the change of basis). We shall see that the Fundamental Theorem in this situation implies (under the assumption that the field F contains all the "eigenvalues" for the given linear transformation T) that there is a basis for V so that the associated matrix for T is *as close to being a diagonal matrix as possible* and so has a particularly simple form. This is the *Jordan canonical form*. The *rational canonical form* is another simple form for the matrix for T (that does not require the eigenvalues for T to be elements of F). In this way we shall be able to give canonical forms for arbitrary $n \times n$ matrices over fields F, that is, find matrices which are similar to a given $n \times n$ matrix and which are particularly simple (almost diagonal, for example).

Example

Let $V = \mathbb{Q}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{Q}\}$ be the usual 3-dimensional vector space of ordered 3-tuples with entries from the field $F = \mathbb{Q}$ of rational numbers and suppose T is the linear transformation

$$T(x, y, z) = (9x + 4y + 5z, -4x - 3z, -6x - 4y - 2z), \qquad x, y, z \in \mathbb{Q}.$$

If we take the standard basis $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$ for V then the matrix A representing this linear transformation is

$$A = \begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}.$$

We shall see that the Jordan canonical form for this matrix A is the much simpler matrix

$$B = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

obtained by taking instead the basis $f_1 = (2, -1, -2)$, $f_2 = (1, 0, -1)$, $f_3 = (3, -2, -2)$ for V:

$$\begin{aligned} T(f_1) &= T(2, -1, -2) = (4, -2, -4) = 2 \cdot f_1 + 0 \cdot f_2 + 0 \cdot f_3 \\ T(f_2) &= T(1, 0, -1) = (4, -1, -4) = 1 \cdot f_1 + 2 \cdot f_2 + 0 \cdot f_3 \\ T(f_3) &= T(3, -2, -2) = (9, -6, -6) = 0 \cdot f_1 + 0 \cdot f_2 + 3 \cdot f_3, \end{aligned}$$

so the columns of the matrix representing T with respect to this basis are $(2, 0, 0)$, $(1, 2, 0)$ and $(0, 0, 3)$, i.e. T has matrix B with respect to this basis. In particular A is similar to the simpler matrix B.

In fact this linear transformation T *cannot* be diagonalized (i.e. there is no choice of basis for V for which the corresponding matrix is a diagonal matrix) so that the matrix B is as close to a diagonal matrix for T as is possible.

The first section below gives some general definitions and states and proves the Fundamental Theorem over an arbitrary P.I.D., after which we return to the application to canonical forms (the application to abelian groups appears in Chapter 5).

These applications can be read independently of the general proof. An alternate and computationally useful proof valid for Euclidean Domains (so in particular for the rings $\mathbb{Z}$ and $F[x]$) along the lines of row and column operations is outlined in the exercises.

12.1 THE BASIC THEORY

We first describe some general finiteness conditions. Let R be a ring and let M be a left R-module.

Definition

(1) The left R-module M is said to be a *Noetherian R–module* or to satisfy the *ascending chain condition on submodules* (or *A.C.C. on submodules*) if there are no infinite increasing chains of submodules, i.e. whenever

$$M_1 \subseteq M_2 \subseteq M_3 \subseteq \cdots$$

is an increasing chain of submodules of M, then there is a positive integer m such that for all $k \geq m$, $M_k = M_m$ (so the chain becomes stationary at stage m: $M_m = M_{m+1} = M_{m+2} = \ldots$).

(2) The ring R is said to be *Noetherian* if it is Noetherian as a left module over itself, i.e. if there are no infinite increasing chains of left ideals in R.

One can formulate analogous notions of A.C.C. on right and on two-sided ideals in a (possibly noncommutative) ring R. For noncommutative rings these properties need not be related.

Theorem 1. Let R be a ring and let M be a left R-module. Then the following are equivalent:

(1) M is a Noetherian R-module.

(2) Every nonempty set of submodules of M contains a maximal element under inclusion.

(3) Every submodule of M is finitely generated.

Proof: [(1) implies (2)] Assume M is Noetherian and let Σ be any nonempty collection of submodules of M. Choose any $M_1 \in \Sigma$. If M_1 is a maximal element of Σ, (2) holds, so assume M_1 is not maximal. Then there is some $M_2 \in \Sigma$ such that $M_1 \subset M_2$. If M_2 is maximal in Σ, (2) holds, so we may assume there is an $M_3 \in \Sigma$ properly containing M_2. Proceeding in this way one sees that if (2) fails we can produce an infinite strictly increasing chain of elements of Σ, contrary to (1).

[(2) implies (3)] Assume (2) holds and let N be any submodule of M. Let Σ be the collection of all finitely generated submodules of N. Since $\{0\} \in \Sigma$, this collection is nonempty. By (2) Σ contains a maximal element N'. If $N' \neq N$, let $x \in N - N'$. Since $N' \in \Sigma$, the submodule N' is finitely generated by assumption, hence also the submodule generated by N' and x is finitely generated. This contradicts the maximality of N', so $N = N'$ is finitely generated.

[(3) implies (1)] Assume (3) holds and let $M_1 \subseteq M_2 \subseteq M_3 \ldots$ be a chain of submodules of M. Let

$$N = \bigcup_{i=1}^{\infty} M_i.$$

By (3) N is finitely generated by, say, $a_1, a_2, \ldots, a_n$. Since $a_i \in N$ for all i, each a_i lies in one of the submodules in the chain, say M_{j_i}. Let $m = \max\{j_1, j_2, \ldots, j_n\}$. Then $a_i \in M_m$ for all i so the module they generate is contained in M_m, i.e. $N \subseteq M_m$. This implies $M_m = N = M_k$ for all $k \geq m$, which proves (1).

Corollary 2. If R is a P.I.D. then every nonempty set of ideals of R has a maximal element and R is a Noetherian ring.

Proof: The P.I.D. R satisfies condition (3) in the theorem with $M = R$.

Recall that even if M itself is a finitely generated R-module, submodules of M need not be finitely generated, so the condition that M be a Noetherian R-module is in general stronger than the condition that M be a finitely generated R-module.

We require a result on "linear dependence" before turning to the main results of this chapter.

Proposition 3. Let R be an integral domain and let M be a free R-module of rank $n < \infty$. Then any $n + 1$ elements of M are R-linearly dependent, i.e. for any $y_1, y_2, \ldots, y_{n+1} \in M$ there are elements $r_1, r_2, \ldots, r_{n+1} \in R$, not all zero, such that

$$r_1 y_1 + r_2 y_2 + \ldots + r_{n+1} y_{n+1} = 0.$$

Proof: The quickest way of proving this is to embed R in its quotient field F (since R is an integral domain) and observe that since $M \cong R \oplus R \oplus \cdots \oplus R$ (n times) we obtain $M \subseteq F \oplus F \oplus \cdots \oplus F$. The latter is an n-dimensional vector space over F so any $n + 1$ elements of M are F-linearly dependent. By clearing the denominators of the scalars (by multiplying through by the product of all the denominators, for example), we obtain an R-linear dependence relation among the $n + 1$ elements of M.

Alternatively, let $e_1, e_2, \ldots, e_n$ be a basis of the free R-module M and let $y_1, y_2, \ldots, y_{n+1}$ be any $n + 1$ elements of M. For $1 \leq i \leq n + 1$ write $y_i = a_{1i}e_i + a_{2i}e_2 + \ldots + a_{ni}e_i$ in terms of the basis $e_1, e_2, \ldots, e_n$. Let A be the $(n+1) \times (n+1)$ matrix whose i, j entry is a_{ij}, $1 \leq i \leq n$, $1 \leq j \leq n + 1$ and whose last row is zero, so certainly $\det A = 0$. Since R is an integral domain, Corollary 11.24 shows that the columns of A are R-linearly dependent. Any dependence relation on the columns of A gives a dependence relation on the y_i's, completing the proof.

If R is any integral domain and M is any R-module recall that

$$\text{Tor}\,(M) = \{x \in M \mid rx = 0 \text{ for some nonzero } r \in R\}$$

is a submodule of M (called *the* torsion submodule of M) and if N is any submodule of Tor (M), N is called *a* torsion submodule of M (so the torsion submodule of M

is the union of all torsion submodules of M, i.e. is the maximal torsion submodule of M). If Tor $(M) = 0$, the module M is said to be *torsion free*.

For any submodule N of M, the *annihilator* of N is the ideal of R defined by

$$\text{Ann}(N) = \{r \in R \mid rn = 0 \text{ for all } n \in N\}.$$

Note that if N is not a torsion submodule of M then $\text{Ann}(N) = (0)$. It is easy to see that if N, L are submodules of M with $N \subseteq L$, then $\text{Ann}(L) \subseteq \text{Ann}(N)$. If R is a P.I.D. and $N \subseteq L \subseteq M$ with $\text{Ann}(N) = (a)$ and $\text{Ann}(L) = (b)$, then $a \mid b$. In particular, the annihilator of any element x of M divides the annihilator of M (this is implied by Lagrange's Theorem when $R = \mathbb{Z}$).

Definition. For any integral domain R the *rank* of an R-module M is the maximum number of R-linearly independent elements of M.

The preceding proposition states that for a free R-module M over an integral domain the rank of a submodule is bounded by the rank of M. This notion of rank agrees with previous uses of the same term. If the ring $R = F$ is a field, then the rank of an R-module M is the dimension of M as a vector space over F and any maximal set of F-linearly independent elements is a basis for M. For a general integral domain, however, an R-module M of rank n need not have a 'basis', i.e. need not be a *free* R-module even if M is torsion free, so some care is necessary with the notion of rank, particularly with respect to the torsion elements of M. The first few exercises following Corollary 11 give an alternate characterization of the rank and provide some examples of (torsion free) R-modules (of rank 1) which are not free.

The next important result shows that if N is a submodule of a free module of finite rank over a P.I.D. then N is again a free module of finite rank and furthermore it is possible to choose generators for the two modules which are related in a simple way.

Theorem 4. Let R be a Principal Ideal Domain, let M be a free R-module of finite rank n and let N be a submodule of M. Then

(1) N is free of rank m, $m \leq n$ and

(2) there exists a basis $y_1, y_2, \ldots, y_n$ of M so that $a_1y_1, a_2y_2, \ldots, a_my_m$ is a basis of N where $a_1, a_2, \ldots, a_m$ are nonzero elements of R with the divisibility relations

$$a_1 \mid a_2 \mid \cdots \mid a_m.$$

Proof: The theorem is trivial for $N = \{0\}$, so assume $N \neq \{0\}$. For each R-module homomorphism φ of M into R, the image $\varphi(N)$ of N is a submodule of R, i.e. an ideal in R. Since R is a P.I.D. this ideal must be principal, say $\varphi(N) = (a_\varphi)$, for some $a_\varphi \in R$. Let

$$\Sigma = \{(a_\varphi) \mid \varphi \in \text{Hom}_R(M, R)\}$$

be the collection of the principal ideals in R obtained in this way from the R-module homomorphisms of M into R. The collection Σ is certainly nonempty since taking φ to be the trivial homomorphism shows that $(0) \in \Sigma$. By Corollary 2, Σ has at least one maximal element i.e. there is at least one homomorphism ν of M to R so that

the principal ideal $\nu(N) = (a_\nu)$ is not properly contained in any other element of Σ. Let $a_1 = a_\nu$ for this maximal element and let $y \in N$ be an element mapping to the generator a_1 under the homomorphism ν: $\nu(y) = a_1$.

We now show the element a_1 is nonzero. Let $x_1, x_2, \ldots, x_n$ be any basis of the free module M and let $\pi_i \in \mathrm{Hom}_R(M, R)$ be the natural projection homomorphism onto the ith coordinate with respect to this basis. Since $N \neq \{0\}$, there exists an i such that $\pi_i(N) \neq 0$, which in particular shows that Σ contains more than just the trivial ideal (0). Since (a_1) is a maximal element of Σ it follows that $a_1 \neq 0$.

We next show that this element a_1 divides $\varphi(y)$ for every $\varphi \in \mathrm{Hom}_R(M, R)$. To see this let d be a generator for the principal ideal generated by a_1 and $\varphi(y)$. Then d is a divisor of both a_1 and $\varphi(y)$ in R and $d = r_1 a_1 + r_2\varphi(y)$ for some $r_1, r_2 \in R$. Consider the homomorphism $\psi = r_1\nu + r_2\varphi$ from M to R. Then $\psi(y) = (r_1\nu + r_2\varphi)(y) = r_1 a_1 + r_2\varphi(y) = d$ so that $d \in \psi(N)$, hence also $(d) \subseteq \psi(N)$. But d is a divisor of a_1 so we also have $(a_1) \subseteq (d)$. Then $(a_1) \subseteq (d) \subseteq \psi(N)$ and by the maximality of (a_1) we must have equality: $(a_1) = (d) = \psi(N)$. In particular $(a_1) = (d)$ shows that $a_1 \mid \varphi(y)$ since d divides $\varphi(y)$.

If we apply this to the projection homomorphisms π_i we see that a_1 divides $\pi_i(y)$ for all i. Write $\pi_i(y) = a_1 b_i$ for some $b_i \in R$, $1 \leq i \leq n$ and define

$$y_1 = \sum_{i=1}^{n} b_i x_i.$$

Note that $a_1 y_1 = y$. Since $a_1 = \nu(y) = \nu(a_1 y_1) = a_1\nu(y_1)$ and a_1 is a nonzero element of the integral domain R this shows

$$\nu(y_1) = 1.$$

We now verify that this element y_1 can be taken as one element in a basis for M and that $a_1 y_1$ can be taken as one element in a basis for N, namely that we have:

(a) $M = Ry_1 \oplus \ker \nu$, and

(b) $N = Ra_1 y_1 \oplus (N \cap \ker \nu)$.

To see (a) let x be an arbitrary element in M and write $x = \nu(x)y_1 + (x - \nu(x)y_1)$. Since

$$\begin{aligned} \nu(x - \nu(x)y_1) &= \nu(x) - \nu(x)\nu(y_1) \\ &= \nu(x) - \nu(x) \cdot 1 \\ &= 0 \end{aligned}$$

we see that $x - \nu(x)y_1$ is an element in the kernel of ν. This shows that x can be written as the sum of an element in Ry_1 and an element in the kernel of ν, so $M = Ry_1 + \ker \nu$. To see that the sum is direct, suppose ry_1 is also an element in the kernel of ν. Then $0 = \nu(ry_1) = r\nu(y_1) = r$ shows that this element is indeed 0.

For (b) observe that $\nu(x')$ is divisible by a_1 for every $x' \in N$ by the definition of a_1 as a generator for $\nu(N)$. If we write $\nu(x') = ba_1$ where $b \in R$ then the decomposition we used in (a) above is $x' = \nu(x)y_1 + (x - \nu(x)y_1) = ba_1 y_1 + (x' - ba_1 y_1)$ where the second summand is in the kernel of ν and is an element of N. This shows that $N = Ra_1 y_1 + (N \cap \ker \nu)$. The fact that the sum in (b) is direct is a special case of the directness of the sum in (a).

We now prove part (1) of the theorem by induction on the rank, m, of N. If $m = 0$, $N = 0$ so (1) holds trivially. Assume then that $m > 0$. Since the sum in (b) above is direct we see easily that $N \cap \ker \nu$ has rank $m - 1$ (cf. the exercises). By induction $N \cap \ker \nu$ is then a free R-module of rank $m - 1$. Again by the directness of the sum in (b) we see that adjoining $a_1 y_1$ to any basis of $N \cap \ker \nu$ gives a basis of N, so N is also free (of rank m), which proves (1).

Finally, we prove (2) by induction on n, the rank of M. Applying (1) to the submodule $\ker \nu$ shows that this submodule is free and because the sum in (a) is direct it is free of rank $n - 1$. By the induction assumption applied to the module $\ker \nu$ (which plays the role of M) and its submodule $\ker \nu \cap N$ (which plays the role of N), we see that there is a basis $y_2, y_3, \ldots, y_n$ of $\ker \nu$ such that $a_2 y_2, a_3 y_3, \ldots, a_m y_m$ is a basis of $N \cap \ker \nu$ for some elements $a_2, a_3, \ldots, a_m$ of R with $a_2 \mid a_3 \mid \cdots \mid a_m$. Since the sums (a) and (b) are direct, $y_1, y_2, \ldots, y_n$ is a basis of M and $a_1 y_1, a_2 y_2, \ldots, a_m y_m$ is a basis of N. To complete the induction it remains to show that a_1 divides a_2. Define a homomorphism φ from M to R by defining $\varphi(y_1) = \varphi(y_2) = 1$ and $\varphi(y_i) = 0$, for all $i > 2$, on the basis for M. Then for this homomorphism φ we have $a_1 = \varphi(a_1 y_1)$ so $a_1 \in \varphi(N)$ hence also $(a_1) \subseteq \varphi(N)$. By the maximality of (a_1) in Σ it follows that $(a_1) = \varphi(N)$. Since $a_2 = \varphi(a_2 y_2) \in \varphi(N)$ we then have $a_2 \in (a_1)$ i.e. $a_1 \mid a_2$. This completes the proof of the theorem.

Recall that the left R-module C is a *cyclic* R-module (for any ring R, not necessarily commutative nor with 1) if there is an element $x \in C$ such that $C = Rx$. We can then define an R-module homomorphism

$$\pi : R \to C$$

by $\pi(r) = rx$, which will be surjective by the assumption $C = Rx$. The First Isomorphism Theorem gives an isomorphism of (left) R-modules

$$R/\ker \pi \cong C.$$

If R is a P.I.D., $\ker \pi$ is a principal ideal, (a), so we see that the cyclic R-modules C are of the form $R/(a)$ where $(a) = \mathrm{Ann}(C)$.

The cyclic modules are the simplest modules (since they require only one generator). The existence portion of the Fundamental Theorem states that any finitely generated module over a P.I.D. is isomorphic to the direct sum of finitely many cyclic modules.

Theorem 5. (Fundamental Theorem, Existence: Invariant Factor Form) Let R be a P.I.D. and let M be a finitely generated R-module.

(1) Then M is isomorphic to the direct sum of finitely many cyclic modules. More precisely,

$$M \cong R^r \oplus R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m)$$

for some integer $r \geq 0$ and nonzero elements $a_1, a_2, \ldots, a_m$ of R which are not units in R and which satisfy the divisibility relations

$$a_1 \mid a_2 \mid \cdots \mid a_m.$$

(2) M is torsion free if and only if M is free.

(3) In the decomposition in (1),

$$\text{Tor}\,(M) \cong R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m).$$

In particular M is a torsion module if and only if $r = 0$ and in this case the annihilator of M is the ideal (a_m).

Proof: The module M can be generated by a finite set of elements by assumption so let $x_1, x_2, \ldots, x_n$ be a set of generators of M of minimal cardinality. Let R^n be the free R-module of rank n with basis $b_1, b_2, \ldots, b_n$ and define the homomorphism $\pi : R^n \to M$ by defining $\pi(b_i) = x_i$ for all i, which is automatically surjective since $x_1, \ldots, x_n$ generate M. By the First Isomorphism Theorem for modules we have $R^n/\ker \pi \cong M$. Now, by Theorem 4 applied to R^n and the submodule $\ker \pi$ we can choose another basis $y_1, y_2, \ldots, y_n$ of R^n so that $a_1y_1, a_2y_2, \ldots, a_my_m$ is a basis of $\ker \pi$ for some elements $a_1, a_2, \ldots, a_m$ of R with $a_1 \mid a_2 \mid \cdots \mid a_m$. This implies

$$M \cong R^n/\ker \pi = (Ry_1 \oplus Ry_2 \oplus \cdots \oplus Ry_n)/(Ra_1y_1 \oplus Ra_2y_2 \oplus \cdots \oplus Ra_my_m).$$

To identify the quotient on the right hand side we use the natural surjective R-module homomorphism

$$Ry_1 \oplus Ry_2 \oplus \cdots \oplus Ry_n \to R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m) \oplus R^{n-m}$$

defined by mapping $(\alpha_1y_1, \ldots, \alpha_ny_n)$ to $(\alpha_1 \bmod (a_1), \ldots, \alpha_m \bmod (a_m), \alpha_{m+1}, \ldots, \alpha_n$ The kernel of this map is clearly the set of elements where a_i divides α_i, $i = 1, 2, \ldots, m$, i.e. $Ra_1y_1 \oplus Ra_2y_2 \oplus \cdots \oplus Ra_my_m$ (cf. Exercise 7). Hence we obtain

$$M \cong R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m) \oplus R^{n-m}.$$

If a is a unit in R then $R/(a) = 0$, so in this direct sum we may remove any of the initial a_i which are units. This gives the decomposition in (1) (with $r = n - m$).

Since $R/(a)$ is a torsion R-module for any nonzero element a of R, (1) immediately implies M is a torsion free module if and only if $M \cong R^r$, which is (2). Part (3) is immediate from the definitions since the annihilator of $R/(a)$ is evidently the ideal (a).

We shall shortly prove the uniqueness of the decomposition in Theorem 5, namely that if we have

$$M \cong R^{r'} \oplus R/(b_1) \oplus R/(b_2) \oplus \cdots \oplus R/(b_{m'})$$

for some integer $r' \geq 0$ and nonzero elements $b_1, b_2, \ldots, b_{m'}$ of R which are not units with

$$b_1 \mid b_2 \mid \cdots \mid b_{m'},$$

then $r = r'$, $m = m'$ and $(a_i) = (b_i)$ (so $a_i = b_i$ up to units) for all i. It is precisely the divisibility condition $a_1 \mid a_2 \mid \cdots \mid a_m$ which gives this uniqueness.

Definition. The integer r in Theorem 5 is called the *free rank* or the *Betti number* of M and the elements $a_1, a_2, \ldots, a_m \in R$ (defined up to multiplication by units in R) are called the *invariant factors* of M.

Note that until we have proved that the invariant factors of M are unique we should properly refer to *a* set of invariant factors for M (and similarly for the free rank), by which we mean any elements giving a decomposition for M as in (1) of the theorem above.

Using the Chinese Remainder Theorem it is possible to decompose the cyclic modules in Theorem 5 further so that M is the direct sum of cyclic modules whose annihilators are as simple as possible (namely (0) or generated by powers of primes in R). This gives an alternate decomposition which we shall also see is unique and which we now describe.

Suppose a is a nonzero element of the Principal Ideal Domain R. Then since R is also a Unique Factorization Domain we can write

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_s^{\alpha_s}$$

where the p_i are distinct primes in R. This factorization is unique up to units, so the ideals $(p_i^{\alpha_i})$, $i = 1, \dots, s$ are uniquely defined. For $i \neq j$ we have $(p_i^{\alpha_i}) + (p_j^{\alpha_j}) = R$ since the sum of these two ideals is generated by a greatest common divisor, which is 1 for distinct primes p_i, p_j. Put another way, the ideals $(p_i^{\alpha_i})$, $i = 1, \dots, s$, are comaximal in pairs. The intersection of all these ideals is the ideal (a) since a is the least common multiple of $p_1^{\alpha_1}, p_2^{\alpha_2}, \dots, p_s^{\alpha_s}$. Then the Chinese Remainder Theorem (Theorem 7.15) shows that

$$R/(a) \cong R/(p_1^{\alpha_1}) \oplus R/(p_2^{\alpha_2}) \oplus \cdots \oplus R/(p_s^{\alpha_s})$$

as rings and also as R-modules.

Applying this to the modules in Theorem 5 allows us to write each of the direct summands $R/(a_i)$ for the invariant factor a_i of M as a direct sum of cyclic modules whose annihilators are the prime power divisors of a_i. This proves:

Theorem 6. (Fundamental Theorem, Existence: Elementary Divisor Form) Let R be a P.I.D. and let M be a finitely generated R-module. Then M is the direct sum of a finite number of cyclic modules whose annihilators are either (0) or generated by powers of primes in R, i.e.

$$M \cong R^r \oplus R/(p_1^{\alpha_1}) \oplus R/(p_2^{\alpha_2}) \oplus \cdots \oplus R/(p_t^{\alpha_t})$$

where $r \geq 0$ is an integer and $p_1^{\alpha_1}, \dots, p_t^{\alpha_t}$ are positive powers of (not necessarily distinct) primes in R.

We proved Theorem 6 by using the prime power factors of the invariant factors for M. In fact we shall see that the decomposition of M into a direct sum of cyclic modules whose annihilators are (0) or prime powers as in Theorem 6 is unique, i.e. the integer r and the ideals $(p_1^{\alpha_1}), \dots, (p_t^{\alpha_t})$ are uniquely defined for M. These prime powers are given a name:

Definition. Let R be a P.I.D. and let M be a finitely generated R-module as in Theorem 6. The prime powers $p_1^{\alpha_1}, \dots, p_t^{\alpha_t}$ (defined up to multiplication by units in R) are called the *elementary divisors* of M.

Suppose M is a finitely generated torsion module over the Principal Ideal Domain R. If for the *distinct* primes $p_1, p_2, \ldots, p_n$ occurring in the decomposition in Theorem 6 we group together all the cyclic factors corresponding to the same prime p_i we see in particular that M can be written as a direct sum

$$M = N_1 \oplus N_2 \oplus \cdots \oplus N_n$$

where N_i consists of all the elements of M which are annihilated by some power of the prime p_i. This result holds also for modules over R which may not be finitely generated:

Theorem 7. (The Primary Decomposition Theorem) Let R be a P.I.D. and let M be a nonzero torsion R-module (not necessarily finitely generated) with nonzero annihilator a. Suppose the factorization of a into distinct prime powers in R is

$$a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}$$

and let $N_i = \{x \in M \mid p_i^{\alpha_i} x = 0\}$, $1 \leq i \leq n$. Then N_i is a submodule of M with annihilator $p_i^{\alpha_i}$ and is the submodule of M of all elements annihilated by some power of p_i. We have

$$M = N_1 \oplus N_2 \oplus \cdots \oplus N_n.$$

If M is finitely generated then each N_i is the direct sum of finitely many cyclic modules whose annihilators are divisors of $p_i^{\alpha_i}$.

Proof: We have already proved these results in the case where M is finitely generated over R. In the general case it is clear that N_i is a submodule of M with annihilator dividing $p_i^{\alpha_i}$. Since R is a P.I.D. the ideals $(p_i^{\alpha_i})$ and $(p_j^{\alpha_j})$ are comaximal for $i \neq j$, so the direct sum decomposition of M can be proved easily by modifying the argument in the proof of the Chinese Remainder Theorem to apply it to modules. Using this direct sum decomposition it is easy to see that the annihilator of N_i is precisely $p_i^{\alpha_i}$.

Definition. The submodule N_i in the previous theorem is called the *p_i-primary component* of M.

Notice that with this terminology the elementary divisors of a finitely generated module M are just the invariant factors of the primary components of Tor (M).

We now prove the uniqueness of the decompositions in the Fundamental Theorem.

Note that if M is any module over a commutative ring R and a is an element of R then $aM = \{am \mid m \in M\}$ is a submodule of M. Recall also that in a Principal Ideal Domain R the nonzero prime ideals are maximal, hence the quotient of R by a nonzero prime ideal is a field.

Lemma 8. Let R be a P.I.D. and let p be a prime in R. Let F denote the field $R/(p)$.

(1) Let $M = R^r$. Then $M/pM \cong F^r$.

(2) Let $M = R/(a)$ where a is a nonzero element of R. Then

$$M/pM \cong \begin{cases} F & \text{if } p \text{ divides } a \text{ in } R \\ 0 & \text{if } p \text{ does not divide } a \text{ in } R\,. \end{cases}$$

(3) Let $M = R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_k)$ where each a_i is divisible by p. Then $M/pM \cong F^k$.

Proof: (1) There is a natural map from R^r to $(R/(p))^r$ defined by mapping $(\alpha_1, \ldots, \alpha_r)$ to $(\alpha_1 \bmod (p), \ldots, \alpha_r \bmod (p))$. This is clearly a surjective R-module homomorphism with kernel consisting of the r-tuples all of whose coordinates are divisible by p, i.e. pR^r, so $R^r/pR^r \cong (R/(p))^r$, which is (1).

(2) This follows from the Isomorphism Theorems: note first that $p(R/(a))$ is the image of the ideal (p) in the quotient $R/(a)$, hence is $(p)+(a)/(a)$. The ideal $(p)+(a)$ is generated by a greatest common divisor of p and a, hence is (p) if p divides a and is $R = (1)$ otherwise. Hence $pM = (p)/(a)$ if p divides a and is $R/(a) = M$ otherwise. If p divides a then $M/pM = (R/(a))/((p)/(a)) \cong R/(p)$, and if p does not divide a then $M/pM = M/M = 0$, which proves (2).

(3) This follows from (2) as in the proof of part (1) of Theorem 5.

Theorem 9. (Fundamental Theorem, Uniqueness) Let R be a P.I.D.

(1) Two finitely generated R-modules M_1 and M_2 are isomorphic if and only if they have the same free rank and the same list of invariant factors.

(2) Two finitely generated R-modules M_1 and M_2 are isomorphic if and only if they have the same free rank and the same list of elementary divisors.

Proof: If M_1 and M_2 have the same free rank and list of invariant factors or the same free rank and list of elementary divisors then they are clearly isomorphic.

Suppose that M_1 and M_2 are isomorphic. Any isomorphism between M_1 and M_2 maps the torsion in M_1 to the torsion in M_2 so we must have $\mathrm{Tor}\,(M_1) \cong \mathrm{Tor}\,(M_2)$. Then $R^{r_1} \cong M_1/\mathrm{Tor}\,(M_1) \cong M_2/\mathrm{Tor}\,(M_2) \cong R^{r_2}$ where r_1 is the free rank of M_1 and r_2 is the free rank of M_2. Let p be any nonzero prime in R. Then from $R^{r_1} \cong R^{r_2}$ we obtain $R^{r_1}/pR^{r_1} \cong R^{r_2}/pR^{r_2}$. By (1) of the previous lemma, this implies $F^{r_1} \cong F^{r_2}$ where F is the field R/pR. Hence we have an isomorphism of an r_1-dimensional vector space over F with an r_2-dimensional vector space over F, so that $r_1 = r_2$ and M_1 and M_2 have the same free rank.

We are reduced to showing that M_1 and M_2 have the same lists of invariant factors and elementary divisors. To do this we need only work with the isomorphic torsion modules $\mathrm{Tor}\,(M_1)$ and $\mathrm{Tor}\,(M_2)$, i.e. we may as well assume that both M_1 and M_2 are torsion R-modules.

We first show they have the same elementary divisors. It suffices to show that for any fixed prime p the elementary divisors which are a power of p are the same for both M_1 and M_2. If $M_1 \cong M_2$ then the p-primary submodule of M_1 (= the direct sum of the cyclic factors whose elementary divisors are powers of p) is isomorphic to the p-primary submodule of M_2, since these are the submodules of elements which are annihilated by some power of p. We are therefore reduced to the case of proving

that if two modules M_1 and M_2 which have annihilator a power of p are isomorphic then they have the same elementary divisors.

We proceed by induction on the power of p in the annihilator of M_1 (which is the same as the annihilator of M_2 since M_1 and M_2 are isomorphic). If this power is 0, then both M_1 and M_2 are 0 and we are done. Otherwise M_1 (and M_2) have nontrivial elementary divisors. Suppose the elementary divisors of M_1 are given by

$$\text{elementary divisors of } M_1\text{: } \underbrace{p, p, \ldots, p}_{m \text{ times}},\ p^{\alpha_1}, p^{\alpha_2}, \ldots, p^{\alpha_s},$$

where $2 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s$, i.e. M_1 is the direct sum of cyclic modules with generators $x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{m+s}$, say, whose annihilators are $(p), (p), \ldots, (p)$, $(p^{\alpha_1}), \ldots, (p^{\alpha_s})$, respectively. Then the submodule pM_1 has elementary divisors

$$\text{elementary divisors of } pM_1\text{: } p^{\alpha_1 - 1}, p^{\alpha_2 - 1}, \ldots, p^{\alpha_s - 1}$$

since pM_1 is the direct sum of the cyclic modules with generators $px_1, px_2, \ldots, px_m$, $px_{m+1}, \ldots, px_{m+s}$ whose annihilators are $(1), (1), \ldots, (1), (p^{\alpha_1 - 1}), \ldots, (p^{\alpha_s - 1})$, respectively. Similarly, if the elementary divisors of M_2 are given by

$$\text{elementary divisors of } M_2\text{: } \underbrace{p, p, \ldots, p}_{n \text{ times}},\ p^{\beta_1}, p^{\beta_2}, \ldots, p^{\beta_t},$$

where $2 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_t$, then pM_2 has elementary divisors

$$\text{elementary divisors of } pM_2\text{: } p^{\beta_1 - 1}, p^{\beta_2 - 1}, \ldots, p^{\beta_t - 1} \quad .$$

Since $M_1 \cong M_2$, also $pM_1 \cong pM_2$ and the power of p in the annihilator of pM_1 is one less than the power of p in the annihilator of M_1. By induction, the elementary divisors for pM_1 are the same as the elementary divisors for pM_2, i.e. $s = t$ and $\alpha_i - 1 = \beta_i - 1$ for $i = 1, 2, \ldots, s$, hence $\alpha_i = \beta_i$ for $i = 1, 2, \ldots, s$. Finally, since also $M_1/pM_1 \cong M_2/pM_2$ we see from (3) of the lemma above that $F^{m+s} \cong F^{n+t}$, which shows that $m + s = n + t$ hence $m = n$ since we have already seen $s = t$. This proves that the set of elementary divisors for M_1 is the same as the set of elementary divisors for M_2.

We now show that M_1 and M_2 must have the same invariant factors. Suppose $a_1 \mid a_2 \mid \cdots \mid a_m$ are invariant factors for M_1. We obtain a set of elementary divisors for M_1 by taking the prime power factors of these elements. Note that then the divisibility relations on the invariant factors imply that a_m is the product of the largest of the prime powers among these elementary divisors, a_{m-1} is the product of the largest prime powers among these elementary divisors once the factors for a_m have been removed, and so on. If $b_1 \mid b_2 \mid \cdots \mid b_n$ are invariant factors for M_2 then we similarly obtain a set of elementary divisors for M_2 by taking the prime power factors of these elements. But we showed above that the elementary divisors for M_1 and M_2 are the same, and it follows that the same is true of the invariant factors.

Corollary 10. Let R be a P.I.D. and let M be a finitely generated R-module.

(1) The elementary divisors of M are the prime power factors of the invariant factors of M.

(2) The largest invariant factor of M is the product of the largest of the distinct prime powers among the elementary divisors of M, the next largest invariant factor is the product of the largest of the distinct prime powers among the remaining elementary divisors of M, and so on.

Proof: The procedure in (1) gives *a* set of elementary divisors and since the elementary divisors for M are unique by the theorem, it follows that the procedure in (1) gives *the* set of elementary divisors. Similarly for (2).

Corollary 11. The Fundamental Theorem of Finitely Generated Abelian Groups: cf. Theorem 5.3 and Theorem 5.5.

Proof: Take $R = \mathbb{Z}$ in Theorems 5, 6 and 9 (note however that the invariant factors are listed in reverse order in Chapter 5 for computational convenience).

The procedure for passing between elementary divisors and invariant factors in Corollary 10 is described in some detail in Chapter 5 in the case of finitely generated abelian groups.

Note also that if a finitely generated module M is written as a direct sum of cyclic modules of the form $R/(a)$ then the ideals (a) which occur are not in general unique unless some additional conditions are imposed (such as the divisibility condition for the invariant factors or the condition that a be the power of a prime in the case of the elementary divisors). To decide whether two modules are isomorphic it is necessary to first write them in such a standard (or *canonical*) form.

EXERCISES

1. Let M be a module over the integral domain R.
   (a) Suppose x is a nonzero torsion element in M. Show that x and 0 are 'linearly dependent'. Conclude that the rank of Tor (M) is 0, so that in particular any torsion R-module has rank 0.
   (b) Show that the rank of M is the same as the rank of the (torsion free) quotient $M/\text{Tor}\, M$.
2. Let M be a module over the integral domain R.
   (a) Suppose that M has rank n and that $x_1, x_2, \ldots, x_n$ is any maximal set of linearly independent elements of M. Let $N = Rx_1 + \ldots + Rx_n$ be the submodule generated by $x_1, x_2, \ldots, x_n$. Prove that N is isomorphic to R^n and that the quotient M/N is a torsion R-module (equivalently, the elements $x_1, \ldots, x_n$ are linearly independent and for any $y \in M$ there is a nonzero element $r \in R$ such that ry can be written as a linear combination $r_1x_1 + \ldots + r_nx_n$ of the x_i).
   (b) Prove conversely that if M contains a submodule N that is free of rank n (i.e. $N \cong R^n$) such that the quotient M/N is a torsion R-module then M has rank n. [Let $y_1, y_2, \ldots, y_{n+1}$ be any $n+1$ elements of M. Use the fact that M/N is torsion to write r_iy_i as a linear combination of a basis for N for some nonzero elements $r_1, \ldots, r_{n+1}$ of R. Use an argument as in the proof of Proposition 3 to see that the r_iy_i, and hence also the y_i, are linearly dependent.]
3. Let R be an integral domain and let A and B be R-modules of ranks m and n, respectively. Prove that the rank of $A \oplus B$ is $m + n$. [Use the previous exercise].

4. Let R be an integral domain, let M be an R-module and let N be a submodule of M. Suppose M has rank n, N has rank r and the quotient M/N has rank s. Prove that $n = r + s$. [Let $x_1, x_2, \ldots, x_s$ be elements of M whose images in M/N are a maximal set of independent elements and let $x_{s+1}, x_{s+2}, \ldots, x_{s+r}$ be a maximal set of independent elements in N. Prove that $x_1, x_2, \ldots, x_{s+r}$ are linearly independent in M and that for any element $y \in M$ there is a nonzero element $r \in R$ such that ry is a linear combination of these elements. Then use Exercise 2.]
5. Let $R = \mathbb{Z}[x]$ and let $M = (2, x)$ be the ideal generated by 2 and x, considered as a submodule of R. Show that $\{2, x\}$ is not a basis of M. [Find a nontrivial R-linear dependence between these two elements.] Show that the rank of M is 1 but that M is not free of rank 1 (cf. Exercise 2).
6. Show that if R is an integral domain and M is any nonprincipal ideal of R then M is torsion free of rank 1 but is not a free R-module.
7. Let R be any ring, let $A_1, A_2, \ldots, A_m$ be R-modules and let B_i be a submodule of A_i, $1 \le i \le m$. Prove that

$$(A_1 \oplus A_2 \oplus \cdots \oplus A_m)/(B_1 \oplus B_2 \oplus \cdots \oplus B_m) \cong (A_1/B_1) \oplus (A_2/B_2) \oplus \cdots \oplus (A_m/B_m).$$

8. Let R be a P.I.D., let B be a torsion R-module and let p be a prime in R. Prove that if $pb = 0$ for some nonzero $b \in B$, then $\mathrm{Ann}(B) \subseteq (p)$.
9. Give an example of an integral domain R and a nonzero torsion R-module M such that $\mathrm{Ann}(M) = 0$. Prove that if N is a finitely generated torsion R-module then $\mathrm{Ann}(N) \neq 0$.
10. For p a prime in the P.I.D. R and N an R-module prove that the p-primary component of N is a submodule of N and prove that N is the direct sum of its p-primary components (there need not be finitely many of them).
11. Let R be a P.I.D., let a be a nonzero element of R and let $M = R/(a)$. For any prime p of R prove that

$$p^{k-1}M/p^kM \cong \begin{cases} R/(p) & \text{if } k \le n \\ 0 & \text{if } k > n, \end{cases}$$

where n is the power of p dividing a in R.
12. Let R be a P.I.D. and let p be a prime in R.
    (a) Let M be a finitely generated torsion R-module. Use the previous exercise to prove that $p^{k-1}M/p^kM \cong F^{n_k}$ where F is the field $R/(p)$ and n_k is the number of elementary divisors of M which are powers p^α with $\alpha \ge k$.
    (b) Suppose M_1 and M_2 are two finitely generated torsion R-modules. Use (a) to prove that, for every $k \ge 0$, M_1 and M_2 have the same number of elementary divisors p^α with $\alpha \ge k$. Prove that this implies M_1 and M_2 have the same set of elementary divisors.
13. If M is a finitely generated module over the P.I.D. R, describe the structure of $M/\mathrm{Tor}(M)$.
14. Let R be a P.I.D. and let M be a torsion R-module. Prove that M is irreducible (cf. Exercises 9 to 11 of Section 10.3) if and only if $M = Rm$ for any nonzero element $m \in M$ where the annihilator of m is a nonzero prime ideal (p).
15. Prove that if R is a Noetherian ring then R^n is a Noetherian R-module. [Fix a basis of R^n. If M is a submodule of R^n show that the collection of first coordinates of elements of M is a submodule of R hence is finitely generated. Let $m_1, m_2, \ldots, m_k$ be elements of M whose first coordinates generate this submodule of R. Show that any element of M can be written as an R-linear combination of $m_1, m_2, \ldots, m_k$ plus an element of M whose first

coordinate is 0. Prove that $M \cap R^{n-1}$ is a submodule of R^{n-1} where R^{n-1} is the set of elements of R^n with first coordinate 0 and then use induction on n. Also cf. Exercise 11, Section 8.3 and Exercise 8, Section 9.1.]

The following set of exercises outlines a proof of Theorem 5 in the special case where R is a Euclidean Domain using a matrix argument involving row and column operations. This applies in particular to the cases $R = \mathbb{Z}$ and $R = F[x]$ of interest in the applications and is computationally useful.

Let R be a Euclidean Domain and let M be an R-module.

16. Prove that M is finitely generated if and only if there is a surjective R-homomorphism $\varphi : R^n \to M$ for some integer n (this is true for any ring R).

Suppose $\varphi : R^n \to M$ is a surjective R-module homomorphism. By Exercise 15, ker φ is finitely generated. If $x_1, x_2, \ldots, x_n$ is a basis for R^n and $y_1, \ldots, y_m$ are generators for ker φ we have

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \quad i = 1, 2, \ldots, m$$

with coefficients $a_{ij} \in R$. It follows that the homomorphism φ (hence the module structure of M) is determined by the choice of generators for R^n and the matrix $A = (a_{ij})$. Such a matrix A will be called a *relations matrix*.

17. **(a)** Show that interchanging x_i and x_j in the basis for R^n interchanges the ith column with the jth column in the corresponding relations matrix.

(b) Show that, for any $a \in R$, replacing the element x_j by $x_j - ax_i$ in the basis for R^n gives another basis for R^n and that the corresponding relations matrix for this basis is the same as the original relations matrix except that a times the jth column has been added to the ith column. [Note that $\cdots + a_ix_i + \cdots + a_jx_j + \cdots = \cdots + (a_i + aa_j)x_i + \cdots + a_j(x_j - ax_i) + \ldots$.]

18. **(a)** Show that interchanging the generators y_i and y_j interchanges the ith row with the jth row in the relations matrix.

(b) Show that, for any $a \in R$, replacing the element y_j by $y_j - ay_i$ gives another set of generators for ker φ and that the corresponding relations matrix for this choice of generators is the same as the original relations matrix except that $-a$ times the ith row has been added to the jth row.

19. By the previous two exercises we may perform elementary row and column operations on a given relations matrix by choosing different generators for R^n and ker φ. If all relation matrices are the zero matrix then ker $\varphi = 0$ and $M \cong R^n$. Otherwise let a_1 be the (nonzero) g.c.d. (recall R is a Euclidean Domain) of all the entries in a fixed initial relations matrix for M .

(a) Prove that by elementary row and column operations we may assume a_1 occurs in a relations matrix of the form

$$\begin{pmatrix} a_1 & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

where a_1 divides a_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

(b) Prove that there is a relations matrix of the form

$$\begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

where a_1 divides all the entries.

(c) Let a_2 be a g.c.d. of all the entries except the element a_1 in the relations matrix in (b). Prove that there is a relations matrix of the form

$$\begin{pmatrix} a_1 & 0 & 0 & \dots & 0 \\ 0 & a_2 & 0 & \dots & 0 \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

where a_1 divides a_2 and a_2 divides all the other entries of the matrix.

(d) Prove that there is a relations matrix of the form $\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$ where D is a diagonal matrix with nonzero entries $a_1, a_2, \dots, a_k$, $k \leq n$, satisfying

$$a_1 \mid a_2 \mid \cdots \mid a_k.$$

Conclude that

$$M \cong R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_k) \oplus R^{n-k}.$$

If n is not the minimal number of generators required for M then some of the initial elements $a_1, a_2, \dots$ above will be units, so the corresponding direct summands above will be 0. If we remove these irrelevant factors we have produced the invariant factors of the module M. Further, the image of the new generators for R^n corresponding to the direct summands above will then be a set of R-generators for the cyclic submodules of M in its invariant factor decomposition (note that the image in M of the generators corresponding to factors with a_i a unit will be 0). The *column* operations performed in the relations matrix reduction correspond to changing the basis used for R^n as described in Exercise 17:

(a) Interchanging the ith column with the jth column corresponds to interchanging the ith and j^{th} elements in the basis for R^n.

(b) For any $a \in R$, adding a times the jth column to the ith column corresponds to subtracting a times the ith basis element from the jth basis element.

Keeping track of the column operations performed and changing the initial choice of generators for M in the same way therefore gives a set of R-generators for the cyclic submodules of M in its invariant factor decomposition.

This process is quite fast computationally once an initial set of generators for M and initial relations matrix are determined. The element a_1 is determined using the Euclidean Algorithm as the g.c.d. of the elements in the initial relations matrix. Using the row and column operations we can obtain the appropriate linear combination of the entries to produce this g.c.d.

in the (1,1)-position of a new relations matrix. One then subtracts the appropriate multiple of the first column and first row to obtain a matrix as in Exercise 19(b), then iterates this process. Some examples of this procedure in a special case are given at the end of the following section.

12.2 THE RATIONAL CANONICAL FORM

We now apply our results on finitely generated modules in the special case where the P.I.D. is the ring $F[x]$ of polynomials in x with coefficients in a field F.

Let V be a finite dimensional vector space over F of dimension n and let T be a fixed linear transformation of V. As we saw in Chapter 10 we can consider V as an $F[x]$-module where the element x acts on V as the linear transformation T (and so any polynomial in x acts on V as the same polynomial in T). Since V has finite dimension over F by assumption, it is by definition finitely generated as an F-module, hence certainly finitely generated as an $F[x]$-module, so the classification theorems of the preceding section apply.

Any nonzero free $F[x]$-module (being isomorphic to a direct sum of copies of $F[x]$) is an infinite dimensional vector space over F, so if V has finite dimension over F then it must in fact be a torsion $F[x]$-module (i.e. its free rank is 0). It follows from the Fundamental Theorem that then V is isomorphic as an $F[x]$-module to the direct sum of cyclic, torsion $F[x]$-modules. We shall see that this decompositon of V will allow us to choose a basis for V with respect to which the matrix representation for the linear transformation T is in a specific simple form. When we use the invariant factor decomposition of V we obtain the *rational canonical form* for the matrix for T, which we analyze in this section. When we use the elementary divisor decomposition (and when F contains all the eigenvalues of T) we obtain the *Jordan canonical form*, considered in the following section and mentioned earlier as the matrix representing T which is as close to being a diagonal matrix as possible. The uniqueness portion of the Fundamental Theorem ensures that the rational and Jordan canonical forms are unique (which is why they are referred to as *canonical*).

One important use of these canonical forms is to classify the distinct linear transformations of V. In particular they allow us to determine when two matrices represent the same linear transformation, i.e. when two given $n \times n$ matrices are similar.

Note that this will be another instance where the structure of the space being acted upon (the invariant factor decomposition of V for example) is used to obtain significant information on the algebraic objects (in this case the linear transformations) which are acting. This will be considered in the case of *groups* acting on vector spaces in Chapter 15 (and goes under the name of Representation Theory of Groups).

Before describing the rational canonical form in detail we first introduce some linear algebra.

Definition

(1) An element λ of F is called an *eigenvalue* of the linear transformation T if there is a nonzero vector $v \in V$ such that $T(v) = \lambda v$. In this situation v is called an *eigenvector* of T with corresponding eigenvalue λ.

(2) If A is an $n \times n$ matrix with coefficients in F, an element λ is called an *eigenvalue* of A with corresponding eigenvector v if v is a nonzero $n \times 1$ column vector such that $Av = \lambda v$.

(3) If λ is an eigenvalue of the linear transformation T, the set $\{v \in V \mid T(v) = \lambda v\}$ is called the *eigenspace* of T corresponding to the eigenvalue λ. Similarly, if λ is an eigenvalue of the $n \times n$ matrix A, the set of $n \times 1$ matrices v with $Av = \lambda v$ is called the *eigenspace* of A corresponding to the eigenvalue λ.

Note that if we fix a basis $\mathcal{B}$ of V then any linear transformation T of V has an associated $n \times n$ matrix A. Conversely, if A is any $n \times n$ matrix then the map T defined by $T(v) = Av$ for $v \in V$, where the v on the right is the $n \times 1$ vector consisting of the coordinates of v with respect to the fixed basis $\mathcal{B}$ of V, is a linear transformation of V. Then v is an eigenvector of T with corresponding eigenvalue λ if and only if the coordinate vector of v with respect to $\mathcal{B}$ is an eigenvector of A with eigenvalue λ. In other words, the eigenvalues for the linear transformation T are the same as the eigenvalues for the matrix A of T with respect to any fixed basis for V.

Definition. The determinant of a linear transformation from V to V is the determinant of any matrix representing the linear transformation (note that this does not depend on the choice of the basis used).

Proposition 12. The following are equivalent:

(1) λ is an eigenvalue of T

(2) $\lambda I - T$ is a singular linear transformation of V

(3) $\det(\lambda I - T) = 0$.

Proof: Since λ is an eigenvalue of T with corresponding eigenvector v if and only if v is a nonzero vector in the kernel of $\lambda I - T$, it follows that (1) and (2) are equivalent.

(2) and (3) are equivalent by our results on determinants.

Definition. Let x be an indeterminate over F. The polynomial $\det(xI - T)$ is called the *characteristic polynomial* of T and will be denoted $c_T(x)$. If A is an $n \times n$ matrix with coefficients in F, $\det(xI - A)$ is called the *characteristic polynomial* of A and will be denoted $c_A(x)$.

It is easy to see by expanding the determinant that the characteristic polynomial of either T or A is a monic polynomial of degree $n = \dim V$. Proposition 12 says that the set of eigenvalues of T (or A) is precisely the set of roots of the characteristic polynomial of T (of A, respectively). In particular, T has at most n distinct eigenvalues.

We have seen that V considered as a module over $F[x]$ via the linear transformation T is a torsion $F[x]$-module. Let $m(x) \in F[x]$ be the unique monic polynomial generating the annihilator of V in $F[x]$. Equivalently, $m(x)$ is the unique monic polynomial of minimal degree annihilating V (i.e. such that $m(T)$ is the 0 linear transformation), and if $f(x) \in F[x]$ is any polynomial annihilating V, $m(x)$ divides $f(x)$.

Since the ring of all $n \times n$ matrices over F is isomorphic to the collection of all linear transformations of V to itself (an isomorphism is obtained by choosing a basis for V), it follows that for any $n \times n$ matrix A over F there is similarly a unique monic polynomial of minimal degree with $m(A)$ the zero matrix.

Definition. The unique monic polynomial which generates the ideal $\mathrm{Ann}(V)$ in $F[x]$ is called the *minimal polynomial* of T and will be denoted $m_T(x)$. The unique monic polynomial of smallest degree which when evaluated at the matrix A is the zero matrix is called the *minimal polynomial* of A and will be denoted $m_A(x)$.

It is easy to see (cf. Exercise 5) that the degrees of these minimal polynomials are at most n^2 where n is the dimension of V. We shall shortly prove that the minimal polynomial for T is a divisor of the characteristic polynomial for T (this is the *Cayley–Hamilton Theorem*), and similarly for A, so in fact the degrees of these polynomials are at most n.

We now describe the *rational canonical form* of the linear transformation T (respectively, of the $n \times n$ matrix A). By Theorem 5 we have an isomorphism

$$V \cong F[x]/(a_1(x)) \oplus F[x]/(a_2(x)) \oplus \cdots \oplus F[x]/(a_m(x)) \tag{12.1}$$

of $F[x]$-modules where $a_1(x), a_2(x), \ldots, a_m(x)$ are polynomials in $F[x]$ of degree at least one with the divisibility conditions

$$a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x) \quad .$$

These invariant factors $a_i(x)$ are only determined up to a unit in $F[x]$ but since the units of $F[x]$ are precisely the nonzero elements of F (i.e. the nonzero constant polynomials), we may make these polynomials *unique* by stipulating that they be *monic*.

Since the annihilator of V is the ideal $(a_m(x))$ (part (3) of Theorem 5), we immediately obtain:

Proposition 13. The minimal polynomial $m_T(x)$ is the largest invariant factor of V. All the invariant factors of V divide $m_T(x)$.

We shall see below how to calculate not only the minimal polynomial for T but also the other invariant factors.

We now choose a basis for each of the direct summands for V in the decomposition (1) above for which the matrix for T is quite simple. Recall that the linear transformation T acting on the left side of (1) is the element x acting by multiplication on each of the factors on the right side of the isomorphism in (1).

We have seen in the example following Proposition 1 of Chapter 11 that the elements $1, \bar{x}, \bar{x}^2, \ldots, \bar{x}^{k-1}$ give a basis for the vector space $F[x]/(a(x))$ where $a(x) = x^k + b_{k-1}x^{k-1} + \cdots + b_1 x + b_0$ is any monic polynomial in $F[x]$ and $\bar{x} = x \bmod (a(x))$. With respect to this basis the linear transformation of multiplication by x acts in a

simple manner:

$$x: \quad \begin{array}{ll} 1 & \mapsto \bar{x} \\ \bar{x} & \mapsto \bar{x}^2 \\ \bar{x}^2 & \mapsto \bar{x}^3 \\ & \vdots \\ \bar{x}^{k-2} & \mapsto \bar{x}^{k-1} \\ \bar{x}^{k-1} & \mapsto \bar{x}^k = -b_0 - b_1\bar{x} - \cdots - b_{k-1}\bar{x}^{k-1} \end{array}$$

where the last equality is because $a(\bar{x}) = 0$ in $F[x]/(a(x))$, i.e. $\bar{x}^k + b_{k-1}\bar{x}^{k-1} + \cdots + b_1\bar{x} + b_0 = 0$. With respect to this basis, the matrix for multiplication by x is therefore

$$\begin{pmatrix} 0 & 0 & \dots & \dots & \dots & -b_0 \\ 1 & 0 & \dots & \dots & \dots & -b_1 \\ 0 & 1 & \dots & \dots & \dots & -b_2 \\ 0 & 0 & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & \dots & 1 & -b_{k-1} \end{pmatrix}.$$

Such matrices are given a name:

Definition. Let $a(x) = x^k + b_{k-1}x^{k-1} + \cdots + b_1x + b_0$ be any monic polynomial in $F[x]$. The *companion matrix* of $a(x)$ is the $k \times k$ matrix with 1's down the first subdiagonal, $-b_0, -b_1, \dots, -b_{k-1}$ down the last column and zeros elsewhere. The companion matrix of $a(x)$ will be denoted by $\mathcal{C}_{a(x)}$.

We apply this to each of the cyclic modules on the right side of (1) above and let $\mathcal{B}_i$ be the elements of V corresponding to the basis chosen above for the cyclic factor $F[x]/(a_i(x))$ under the isomorphism in (1). Then by definition the linear transformation T acts on $\mathcal{B}_i$ by the companion matrix for $a_i(x)$ since we have seen that this is how multiplication by x acts. The union $\mathcal{B}$ of the $\mathcal{B}_i$'s gives a basis for V since the sum on the right of (1) is direct and with respect to this basis the linear transformation T has as matrix the *direct sum* of the companion matrices for the invariant factors, i.e.

$$\begin{pmatrix} \mathcal{C}_{a_1(x)} & & & \\ & \mathcal{C}_{a_2(x)} & & \\ & & \ddots & \\ & & & \mathcal{C}_{a_m(x)} \end{pmatrix}. \tag{12.2}$$

Notice that this matrix is uniquely determined from the invariant factors of the $F[x]$-module V and, by Theorem 9, the list of invariant factors uniquely determines the module V up to isomorphism as an $F[x]$-module.

Definition

(1) A matrix is said to be in *rational canonical form* if it is the direct sum of companion matrices for polynomials $a_1(x), \dots, a_m(x)$ of degree at least one with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$. The polynomials $a_i(x)$ are called the *invariant fac-*

tors of the matrix. Such a matrix is also said to be a *block diagonal* matrix with blocks the companion matrices for the $a_i(x)$.

(2) A *rational canonical form* for a linear transformation T is a matrix representing T which is in rational canonical form.

We have seen that any linear transformation T has a rational canonical form. We now see that this rational canonical form is unique (hence is called *the* rational canonical form for T). To see this note that the process we used to determine the matrix of T from the direct sum decomposition is reversible. Suppose $b_1(x), b_2(x), \ldots, b_t(x)$ are monic polynomials in $F[x]$ of degree at least one such that $b_i(x) \mid b_{i+1}(x)$ for all i and suppose for some basis $\mathcal{E}$ of V, that the matrix of T with respect to the basis $\mathcal{E}$ is the direct sum of the companion matrices of the $b_i(x)$. Then V must be a direct sum of T-stable subspaces D_i, one for each $b_i(x)$ in such a way that the matrix of T on each D_i is the companion matrix of $b_i(x)$. Let $\mathcal{E}_i$ be the corresponding (ordered) basis of D_i (so $\mathcal{E}$ is the union of the $\mathcal{E}_i$) and let e_i be the first basis element in $\mathcal{E}_i$. Then it is easy to see that D_i is a cyclic $F[x]$-module with generator e_i and that the annihilator of D_i is $b_i(x)$. Thus the torsion $F[x]$-module V decomposes into a direct sum of cyclic $F[x]$-modules in two ways, both of which satisfy the conditions of Theorem 5, i.e. both of which give lists of invariant factors. Since the invariant factors are unique by Theorem 9, $a_i(x)$ and $b_i(x)$ must differ by a unit factor in $F[x]$ and since the polynomials are monic by assumption, we must have $a_i(x) = b_i(x)$ for all i.

We summarize this as:

Theorem 14. (Rational Canonical Form for Linear Transformations) Let V be a finite dimensional vector space over the field F and let T be a linear transformation of V.

(1) There is a basis for V with respect to which the matrix for T is in rational canonical form, i.e. is a block diagonal matrix whose diagonal blocks are the companion matrices for polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ of degree at least one with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$.

(2) The rational canonical form for T is unique.

The use of the word *rational* is to indicate that this canonical form is calculated entirely within the field F and exists for any linear transformation T. This is not the case for the Jordan canonical form (considered later), which only exists if the field F contains the eigenvalues for T (cf. also the remarks following Corollary 18).

The following result translates the notion of similar linear transformations (i.e. the same linear transformation up to a change of basis) into the language of modules and relates this notion to rational canonical forms.

Theorem 15. Let S and T be linear transformations of V. Then the following are equivalent:

(1) S and T are similar linear transformations

(2) the $F[x]$-modules obtained from V via S and via T are isomorphic $F[x]$-modules

(3) S and T have the same rational canonical form.

Proof: [(1) implies (2)] Assume there is a nonsingular linear transformation U such that $S = UTU^{-1}$. The vector space isomorphism $U : V \to V$ is also an $F[x]$-module homomorphism, where x acts on the first V via T and on the second via S, since for example $U(xv) = U(Tv) = UT(v) = SU(v) = x(Uv)$. Hence this is an $F[x]$-module isomorphism of the two modules in (2).

[(2) implies (3)] Assume (2) holds and denote by V_1 the vector space V made into an $F[x]$-module via S and denote by V_2 the space V made into an $F[x]$-module via T. Since $V_1 \cong V_2$ as $F[x]$-modules they have the same list of invariant factors. Thus S and T have a common rational canonical form.

[(3) implies (1)] Assume (3) holds. Since S and T have the same matrix representation with respect to some choice of (possibly different) bases of V by assumption, they are, up to a change of basis, the same linear transformation of V, hence are similar.

Let A be any $n \times n$ matrix with entries from F. Let V be an n-dimensional vector space over F. Recall we can then *define* a linear transformation T on V by choosing a basis for V and setting $T(v) = Av$ where v on the right hand side means the $n \times 1$ column vector of coordinates of v with respect to our chosen basis (this is just the usual identification of linear transformations with matrices). Then (of course) the matrix for this T with respect to this basis is the given matrix A. Put another way, any $n \times n$ matrix A with entries from the field F arises as the matrix for some linear transformation T of an n-dimensional vector space.

This dictionary between linear transformations of vector spaces and matrices allows us to state our previous two results in the language of matrices:

Theorem 16. (Rational Canonical Form for Matrices) Let A be an $n \times n$ matrix over the field F.

(1) The matrix A is similar to a matrix in rational canonical form, i.e. there is an invertible $n \times n$ matrix P over F such that $P^{-1}AP$ is a block diagonal matrix whose diagonal blocks are the companion matrices for polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ of degree at least one with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$.

(2) The rational canonical form for A is unique.

Definition. The *invariant factors* of an $n \times n$ matrix over a field F are the invariant factors of its rational canonical form.

Theorem 17. Let A and B be $n \times n$ matrices over the field F. Then A and B are similar if and only if A and B have the same rational canonical form.

If A is a matrix with entries from a field F and F is a subfield of a larger field K then we may also consider A as a matrix over K. The next result shows that the rational canonical form for A and questions of similarity do not depend on which field contains the entries of A.

Corollary 18. Let A and B be two $n \times n$ matrices over a field F and suppose F is a subfield of the field K.

(1) The rational canonical form of A is the same whether it is computed over K or over F. The minimal and characteristic polynomials and the invariant factors of A are the same whether A is considered as a matrix over F or as a matrix over K.

(2) The matrices A and B are similar over K if and only if they are similar over F, i.e. there exists an invertible $n \times n$ matrix P with entries from K such that $B = P^{-1}AP$ if and only if there exists an (in general different) invertible $n \times n$ matrix Q with entries from F such that $B = Q^{-1}AQ$.

Proof: (1) Let M be the rational canonical form of A when computed over the smaller field F. Since M satisfies the conditions in the definition of the rational canonical form over K, the uniqueness of the rational canonical form implies that M is also the rational canonical form of A over K. Hence the invariant factors of A are the same whether A is viewed over F or over K. In particular, since the minimal polynomial is the largest invariant factor of A it also does not depend on the field over which A is viewed. It is clear from the determinant definition of the characteristic polynomial of A that this polynomial depends only on the entries of A (we shall see shortly that the characteristic polynomial is the product of all the invariant factors for A, which will give an alternate proof of this result).

(2) If A and B are similar over the smaller field F they are clearly similar over K. Conversely, if A and B are similar over K, they have the same rational canonical form over K. By (1) they have the same rational canonical form over F, hence are similar over F by Theorem 17.

This corollary asserts in particular that the rational canonical form for an $n \times n$ matrix A is an $n \times n$ matrix with entries in the smallest field containing the entries of A. Further, this canonical form is the same matrix even if we allow conjugation of A by nonsingular matrices whose entries come from larger fields. This explains the terminology of *rational* canonical form.

The next proposition gives the connection between the characteristic polynomial of a matrix (or of a linear transformation) and its invariant factors and is quite useful for determining these invariant factors (particularly for matrices of small size).

Lemma 19. Let $a(x) \in F[x]$ be any monic polynomial.

(1) The characteristic polynomial of the companion matrix of $a(x)$ is $a(x)$.

(2) If M is the block diagonal matrix which is the direct sum of matrices $A_1, A_2, \ldots, A_k$,

$$M = \begin{pmatrix} A_1 & 0 & \ldots & 0 \\ 0 & A_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_k \end{pmatrix},$$

then the characteristic polynomial of M is the product of the characteristic polynomials of $A_1, A_2, \ldots, A_k$.

Proof: These are both straightforward exercises.

Proposition 20. Let A be an $n \times n$ matrix over the field F.

(1) The characteristic polynomial of A is the product of all the invariant factors of A.

(2) (The Cayley–Hamilton Theorem) The minimal polynomial of A divides the characteristic polynomial of A.

(3) The characteristic polynomial of A divides some power of the minimal polynomial. In particular these polynomials have the same roots, not counting multiplicities.

The same statements are true if the matrix A is replaced by a linear transformation T of an n-dimensional vector space over F.

Proof: Let B be the rational canonical form of A. By the previous lemma the block diagonal form of B shows that the characteristic polynomial of B is the product of the characteristic polynomials of the companion matrices of A. By the first part of the lemma above, the characteristic polynomial of the companion matrix $C_{a(x)}$ for $a(x)$ is just $a(x)$, which implies that the characteristic polynomial for B is the product of the invariant factors of A. Since A and B are similar, they have the same characteristic polynomial, which proves (1). Assertion (2) is immediate from (1) since the minimal polynomial for A is the largest invariant factor of A. The fact that all the invariant factors divide the largest one immediately implies (3). The final assertion is clear from the dictionary between linear transformations of vector spaces and matrices.

Note that part (2) of the proposition is the assertion that the matrix A satisfies its own characteristic polynomial, i.e. $c_A(A) = 0$ as matrices, which is the usual formulation for the Cayley–Hamilton Theorem. Note also that it implies the degree of the minimal polynomial for A has degree at most n, a result mentioned before.

The relations in Proposition 20 are frequently quite useful in the determination of the invariant factors for a matrix A, particularly for matrices of small degree (cf. Exercises 3 and 4 and the examples). The following result (which relies on the exercises at the end of the previous section and whose proof we outline in the exercises) computes the invariant factors in general.

Let A be an $n \times n$ matrix over the field F. Then $xI - A$ is an $n \times n$ matrix with entries in $F[x]$. The three operations:

(a) interchanging two rows or columns

(b) adding a multiple (in $F[x]$) of one row or column to another

(c) multiplying any row or column by a unit in $F[x]$, i.e. by a nonzero element in F,

are called *elementary row and column operations*.

Theorem 21. Let A be an $n \times n$ matrix over the field F. Using the three elementary row and column operations above, the $n \times n$ matrix $xI - A$ with entries from

$F[x]$ can be put into the diagonal form (called the *Smith Normal Form* for A)

$$\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & a_1(x) & & & \\ & & & & a_2(x) & & \\ & & & & & \ddots & \\ & & & & & & a_m(x) \end{pmatrix}$$

with monic nonzero elements $a_1(x), a_2(x), \ldots, a_m(x)$ of $F[x]$ with degrees at least one and satisfying $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$. The elements $a_1(x), \ldots, a_m(x)$ are the invariant factors of A .

Proof: cf. the exercises.

Invariant Factor Decomposition Algorithm: Converting to Rational Canonical Form

As mentioned in the exercises at the end of the previous section, keeping track of the operations necessary to diagonalize $xI - A$ will explicitly give a matrix P such that $P^{-1}AP$ is in rational canonical form. Equivalently, if V is a given $F[x]$–module with vector space basis $[e_1, e_2, \ldots, e_n]$, then P defines the change of basis giving the Invariant Factor Decomposition of V into a direct sum of cyclic $F[x]$–modules. In particular, if A is the matrix of the linear transformation T of the $F[x]$–module V defined by x (i.e. $T(e_j) = xe_j = \sum_{i=1}^{n} a_{ij}e_i$ where $A = (a_{ij})$), then the matrix P defines the change of basis for V with respect to which the matrix for T is in rational canonical form.

We first describe the algorithm in the general context of determining the Invariant Factor Decomposition of a given $F[x]$–module V with vector space basis $[e_1, e_2, \ldots, e_n]$ (the proof is outlined in the exercises). We then describe the algorithm to convert a given $n \times n$ matrix A to rational canonical form (in which reference to an underlying vector space and associated linear transformation are suppressed).

Explicit numerical examples of this algorithm are given in Examples 2 and 3 following.

Invariant Factor Decomposition Algorithm. Let V be an $F[x]$–module with vector space basis $[e_1, e_2, \ldots, e_n]$ (so in particular these elements are generators for V as an $F[x]$–module). Let T be the linear transformation of V to itself defined by x and let A be the $n \times n$ matrix associated to T and this choice of basis for V, i.e.

$$T(e_j) = xe_j = \sum_{i=1}^{n} a_{ij}e_i \qquad \text{where} \qquad A = (a_{ij}).$$

(1) Use the following three elementary row and column operations to diagonalize the matrix $xI - A$ over $F[x]$, keeping track of the *row* operations used:

(a) interchange two rows or columns (denoted $R_i \leftrightarrow R_j$ for the interchange of the ith and j^{th} rows and $C_i \leftrightarrow C_j$ for columns),
(b) add a multiple (in $F[x]$) of one row or column to another (denoted $R_i + p(x)R_j \mapsto R_i$ if $p(x)$ times the jth row is added to the ith row, and similarly $C_i + p(x)C_j \mapsto C_i$ for columns),
(c) multiply any row or column by a unit in $F[x]$, i.e. by a nonzero element in F (denoted uR_i if the ith row is multiplied by $u \in F^\times$, and similarly uC_i for columns).

(2) Beginning with the $F[x]$–module generators $[e_1, e_2, \ldots, e_n]$, for each row operation used in (1), change the set of generators by the following rules:
(a) If the ith row is interchanged with the jth row then interchange the ith and j^{th} generators.
(b) If $p(x)$ times the jth row is added to the ith row then subtract $p(x)$ times the ith generator from the jth generator (note the indices).
(c) If the ith row is multiplied by the unit $u \in F$ then divide the ith generator by u.

(3) When $xI - A$ has been diagonalized to the form in Theorem 21 the generators $[e_1, e_2, \ldots, e_n]$ for V will be in the form of $F[x]$–linear combinations of $e_1, e_2, \ldots, e_n$. Use $xe_j = T(e_j) = \sum_{i=1}^n a_{ij}e_i$ to write these elements as F–linear combinations of $e_1, e_2, \ldots, e_n$. When $xI - A$ has been diagonalized, the first $n-m$ of these linear combinations are 0 (providing a useful numerical check on the computations) and the remaining m linear combinations are nonzero, i.e. the generators for V are in the form $[0, \ldots, 0, f_1, \ldots, f_m]$ corresponding precisely to the diagonal elements in Theorem 21. The elements $f_1, \ldots, f_m$ are a set of $F[x]$-module generators for the cyclic factors in the invariant factor decomposition of V (with annihilators $(a_1(x)), \ldots, (a_m(x))$, respectively):

$$V = F[x]\,f_1 \oplus F[x]\,f_2 \oplus \ldots \oplus F[x]\,f_m\,,$$
$$F[x]\,f_i \cong F[x]/(a_i(x)) \qquad i = 1, 2, \ldots, m\,,$$

giving the Invariant Factor Decomposition of the $F[x]$–module V.

(4) The corresponding *vector space* basis for each cyclic factor of V is then given by the elements $f_i, Tf_i, T^2f_i, \ldots, T^{\deg a_i(x)-1}f_i$.

(5) Write the kth element of the vector space basis computed in (4) in terms of the original vector space basis $[e_1, e_2, \ldots, e_n]$ and use the coordinates for the kth column of an $n \times n$ matrix P. Then $P^{-1}AP$ is in rational canonical form (with diagonal blocks the companion matrices for the $a_i(x)$). This is the matrix for the linear transformation T with respect to the vector space basis in (4).

We now describe the algorithm to convert a given $n \times n$ matrix A to rational canonical form, i.e. to determine an $n \times n$ matrix P so that $P^{-1}AP$ is in rational canonical form. This is nothing more than the algorithm above applied to the vector space $V = F^n$ of $n \times 1$ column vectors with standard basis $[e_1, e_2, \ldots, e_n]$ (where e_i is the column vector with 1 in the ith position and 0's elsewhere) and T is the linear transformation defined by A and this choice of basis. Explicit reference to this

underlying vector space and associated linear transformation are suppressed, so the algorithm is purely matrix theoretic.

Converting an n x n Matrix to Rational Canonical Form. Let A be an $n \times n$ matrix with entries in the field F.

(1) Use the following three elementary row and column operations to diagonalize the matrix $xI - A$ over $F[x]$, keeping track of the *row* operations used:
   (a) interchange two rows or columns (denoted $R_i \leftrightarrow R_j$ for the interchange of the ith and j^{th} rows and $C_i \leftrightarrow C_j$ for columns),
   (b) add a multiple (in $F[x]$) of one row or column to another (denoted $R_i + p(x)R_j \mapsto R_i$ if $p(x)$ times the jth row is added to the ith row, and similarly $C_i + p(x)C_j \mapsto C_i$ for columns),
   (c) multiply any row or column by a unit in $F[x]$, i.e. by a nonzero element in F (denoted uR_i if the ith row is multiplied by $u \in F^\times$, and similarly uC_i for columns).

   Let $d_1, \ldots, d_m$ be the degrees of the monic nonconstant polynomials $a_1(x), \ldots, a_m(x)$ appearing on the diagonal, respectively.

(2) Beginning with the $n \times n$ identity matrix P', for each row operation used in (1), change the matrix P' by the following rules:
   (a) If $R_i \leftrightarrow R_j$ then interchange the ith and j^{th} columns of P' (i.e. $C_i \leftrightarrow C_j$ for P').
   (b) If $R_i + p(x)R_j \mapsto R_i$ then subtract the product of the matrix $p(A)$ times the ith column of P' from the jth column of P' (i.e. $C_j - p(A)C_i \mapsto C_j$ for P' – note the indices).
   (c) If uR_i then divide the elements of the ith column of P' by u (i.e. $u^{-1}C_i$ for P').

(3) When $xI - A$ has been diagonalized to the form in Theorem 21 the first $n - m$ columns of the matrix P' are 0 (providing a useful numerical check on the computations) and the remaining m columns of P' are nonzero. For each $i = 1, 2, \ldots, m$, multiply the ith nonzero column of P' successively by $A^0 = I, A^1, A^2, \ldots, A^{d_i - 1}$, where d_i is the integer in (1) above and use the resulting column vectors (in this order) as the next d_i columns of an $n \times n$ matrix P. Then $P^{-1}AP$ is in rational canonical form (whose diagonal blocks are the companion matrices for the polynomials $a_1(x), \ldots, a_m(x)$ in (1)).

In the theory of canonical forms for linear transformations (or matrices) the characteristic polynomial plays the role of the order of a finite abelian group and the minimal polynomial plays the role of the exponent (after all, they are the same invariants, one for modules over the Principal Ideal Domain $\mathbb{Z}$ and the other for modules over the Principal Ideal Domain $F[x]$) so we can solve problems directly analogous to those we considered for finite abelian groups in Chapter 5. In particular, we can:

(A) determine the rational canonical form of a given matrix (analogous to decomposing a finite abelian group as a direct product of cyclic groups)

(B) determine whether two given matrices are similar (analogous to determining whether two given finite abelian groups are isomorphic)

(C) determine all similarity classes of matrices over F with a given characteristic polynomial (analogous to determining all abelian groups of a given order)

(D) determine all similarity classes of $n \times n$ matrices over F with a given minimal polynomial (analogous to determining all abelian groups of rank at most n of a given exponent).

Examples

(1) We find the rational canonical forms of the following matrices over $\mathbb{Q}$ and determine if they are similar:

$$A = \begin{pmatrix} 2 & -2 & 14 \\ 0 & 3 & -7 \\ 0 & 0 & 2 \end{pmatrix} \quad B = \begin{pmatrix} 0 & -4 & 85 \\ 1 & 4 & -30 \\ 0 & 0 & 3 \end{pmatrix} \quad C = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{pmatrix}.$$

A direct computation shows that all three of these matrices have the same characteristic polynomial: $c_A(x) = c_B(x) = c_C(x) = (x-2)^2(x-3)$. Since the minimal and characteristic polynomials have the same roots, the only possibilities for the minimal polynomials are $(x-2)(x-3)$ or $(x-2)^2(x-3)$. We quickly find that $(A-2I)(A-3I) = 0$, $(B-2I)(B-3I) \neq 0$ (the 1,1-entry is nonzero) and $(C-2I)(C-3I) \neq 0$ (the 1,2-entry is nonzero). It follows that

$$m_A(x) = (x-2)(x-3), \quad m_B(x) = m_C(x) = (x-2)^2(x-3).$$

It follows immediately that there are no additional invariant factors for B and C. Since the invariant factors for A divide the minimal polynomial and have product the characteristic polynomial, we see that A has invariant factors $x-2$, $(x-2)(x-3) = x^2-5x+6$. (For 2×2 and 3×3 matrices the determination of the characteristic and minimal polynomials determines all the invariant factors, cf. Exercises 3 and 4.) We conclude that B and C are similar and neither is similar to A. The rational canonical forms are (note $(x-2)^2(x-3) = x^3-7x^2+16x-12$)

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & -6 \\ 0 & 1 & 5 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 12 \\ 1 & 0 & -16 \\ 0 & 1 & 7 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 12 \\ 1 & 0 & -16 \\ 0 & 1 & 7 \end{pmatrix}.$$

(2) In the example above the rational canonical forms were obtained simply by determining the characteristic and minimal polynomials for the matrices. As mentioned, this is sufficient for 2×2 and 3×3 matrices since this information is sufficient to determine all of the invariant factors. For larger matrices, however, this is in general not sufficient (cf. the next example) and more work is required to determine the invariant factors. In this example we again compute the rational canonical form for the matrix A in Example 1 following the two algorithms outlined above. While this is computationally more difficult for this small matrix (as will be apparent), it has the advantage even in this case that it also explicitly computes a matrix P with $P^{-1}AP$ in rational canonical form.

I. (*Invariant Factor Decomposition*) We use row and column operations (in $\mathbb{Q}[x]$) to reduce the matrix

$$xI - A = \begin{pmatrix} x-2 & 2 & -14 \\ 0 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix}$$

to diagonal form. As in the invariant factor decomposition algorithm, we shall use the notation $R_i \leftrightarrow R_j$ to denote the interchange of the ith and j^{th} rows, $R_i + aR_j \mapsto R_i$ if a times the jth row is added to the ith row, simply uR_i if the ith row is multiplied by u (and similarly for columns, using C instead of R). Note also that the first two operations we perform below are rather *ad hoc* and were chosen simply to have integers everywhere in the computation:

$$\begin{pmatrix} x-2 & 2 & -14 \\ 0 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \underset{\substack{R_1+R_2 \\ \mapsto R_1}}{\longrightarrow} \begin{pmatrix} x-2 & x-1 & -7 \\ 0 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \rightarrow$$

$$\underset{\substack{C_1-C_2 \\ \mapsto C_1}}{\longrightarrow} \begin{pmatrix} -1 & x-1 & -7 \\ -x+3 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \underset{-R_1}{\longrightarrow} \begin{pmatrix} 1 & -x+1 & 7 \\ -x+3 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \rightarrow$$

$$\underset{\substack{R_2+(x-3)R_1 \\ \mapsto R_2}}{\longrightarrow} \begin{pmatrix} 1 & -x+1 & 7 \\ 0 & -x^2+5x-6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \underset{\substack{C_2+(x-1)C_1 \\ \mapsto C_2}}{\longrightarrow} \begin{pmatrix} 1 & 0 & 7 \\ 0 & -x^2+5x-6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix}$$

$$\underset{\substack{C_3-7C_1 \\ \mapsto C_3}}{\longrightarrow} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -x^2+5x-6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \underset{-C_2}{\longrightarrow} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x^2-5x+6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \rightarrow$$

$$\underset{\substack{R_2-7R_3 \\ \mapsto R_2}}{\longrightarrow} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x^2-5x+6 & 0 \\ 0 & 0 & x-2 \end{pmatrix} \underset{\substack{R_2 \leftrightarrow R_3 \\ C_2 \leftrightarrow C_3}}{\longrightarrow} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x-2 & 0 \\ 0 & 0 & x^2-5x+6 \end{pmatrix}.$$

This determines the invariant factors $x-2, x^2-5x+6$ for this matrix, which we determined in Example 1 above. Let now V be a 3-dimensional vector space over $\mathbb{Q}$ with basis e_1, e_2, e_3 and let T be the corresponding linear transformation (which defines the action of x on V), i.e.

$$\begin{aligned} xe_1 &= T(e_1) = 2e_1 \\ xe_2 &= T(e_2) = -2e_1 + 3e_2 \\ xe_3 &= T(e_3) = 14e_1 - 7e_2 + 2e_3. \end{aligned}$$

The row operations used in the reduction above were $R_1 + R_2 \mapsto R_1$, $-R_1$, $R_2 + (x-3)R_1 \mapsto R_2$, $R_2 - 7R_3 \mapsto R_2$, $R_2 \leftrightarrow R_3$. Starting with the basis $[e_1, e_2, e_3]$ for V and changing it according to the rules given in the text, we obtain

$$[e_1, e_2, e_3] \rightarrow [e_1, e_2-e_1, e_3] \rightarrow [-e_1, e_2-e_1, e_3] \rightarrow [-e_1-(x-3)(e_2-e_1), e_2-e_1, e_3]$$
$$\rightarrow [-e_1-(x-3)(e_2-e_1), e_2-e_1, e_3+7(e_2-e_1)] \rightarrow [-e_1-(x-3)(e_2-e_1), e_3+7(e_2-e$$

Using the formulas above for the action of x, we see that these last elements are the elements $[0, -7e_1 + 7e_2 + e_3, -e_1 + e_2]$ of V corresponding to the elements $1, x - 2$ and $x^2 - 5x + 6$ in the diagonalized form of $xI - A$, respectively. The elements $f_1 = -7e_1 + 7e_2 + e_3$ and $f_2 = -e_1 + e_2$ are therefore $\mathbb{Q}[x]$-module generators for the two cyclic factors of V in its invariant factor decomposition as a $\mathbb{Q}[x]$-module. The corresponding $\mathbb{Q}$-vector space bases for these two factors are then f_1 and $f_2, xf_2 = Tf_2$, i.e. $-7e_1 + 7e_2 + e_3$ and $-e_1 + e_2, T(-e_1 + e_2) = -4e_1 + 3e_2$. Then the matrix

$$P = \begin{pmatrix} -7 & -1 & -4 \\ 7 & 1 & 3 \\ 1 & 0 & 0 \end{pmatrix}$$

conjugates A into its rational canonical form:

$$P^{-1}AP = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & -6 \\ 0 & 1 & 5 \end{pmatrix},$$

as one easily checks.

II. (*Converting A directly to Rational Canonical Form*) We use the row operations involved in the diagonalization of $xI - A$ to determine the matrix P' of the algorithm above:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \underset{\substack{C_2 - C_1 \\ \mapsto C_2}}{\longrightarrow} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \underset{-C_1}{\longrightarrow} \begin{pmatrix} -1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \rightarrow$$

$$\underset{\substack{C_1 - (A-3I)C_2 \\ \mapsto C_1}}{\longrightarrow} \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \underset{\substack{C_3 + 7C_2 \\ \mapsto C_3}}{\longrightarrow} \begin{pmatrix} 0 & -1 & -7 \\ 0 & 1 & 7 \\ 0 & 0 & 1 \end{pmatrix} \underset{C_2 \leftrightarrow C_3}{\longrightarrow} \begin{pmatrix} 0 & -7 & -1 \\ 0 & 7 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Here we have $d_1 = 1$ and $d_2 = 2$, corresponding to the second and third nonzero columns of P', respectively. The columns of P are therefore given by

$$\begin{pmatrix} -7 \\ 7 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad A\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -4 \\ 3 \\ 0 \end{pmatrix},$$

respectively, which again gives the matrix P above.

(3) For the 3×3 matrix A it was not necessary to perform the lengthy calculations above merely to determine the rational canonical form (equivalently, the invariant factors), as we saw in Example 1. For $n \times n$ matrices with $n \geq 4$, however, the computation of the characteristic and minimal polynomials is in general not sufficient for the determination of all the invariant factors, so the more extensive calculations of the previous example may become necessary. For example, consider the matrix

$$D = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix}.$$

A short computation shows that the characteristic polynomial of D is $(x-1)^4$. The possible minimal polynomials are then $x - 1$, $(x-1)^2$, $(x-1)^3$ and $(x-1)^4$. Clearly

$D - I \neq 0$ and another short computation shows that $(D-I)^2 = 0$, so the minimal polynomial for D is $(x-1)^2$. There are then two possible sets of invariant factors:

$$x-1, x-1, (x-1)^2 \qquad \text{and} \qquad (x-1)^2, (x-1)^2.$$

To determine the invariant factors for D we apply the procedure of the previous example to the 4×4 matrix

$$xI - D = \begin{pmatrix} x-1 & -2 & 4 & -4 \\ -2 & x+1 & -4 & 8 \\ -1 & 0 & x-1 & 2 \\ 0 & -1 & 2 & x-3 \end{pmatrix}.$$

The diagonal matrix obtained from this matrix by elementary row and column operations is the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & (x-1)^2 & 0 \\ 0 & 0 & 0 & (x-1)^2 \end{pmatrix},$$

which shows that the invariant factors for D are $(x-1)^2, (x-1)^2$ (one series of elementary row and column operations which diagonalize $xI - D$ are: $R_1 \leftrightarrow R_3$, $-R_1$, $R_2 + 2R_1 \mapsto R_2$, $R_3 - (x-1)R_1 \mapsto R_3$, $C_3 + (x-1)C_1 \mapsto C_3$, $C_4 + 2C_1 \mapsto C_4$, $R_2 \leftrightarrow R_4$, $-R_2$, $R_3 + 2R_2 \mapsto R_3$, $R_4 - (x+1)R_2 \mapsto R_4$, $C_3 + 2C_2 \mapsto C_3$, $C_4 + (x-3)C_2 \mapsto C_4$).

I. (*Invariant Factor Decomposition*) If e_1, e_2, e_3, e_4 is a basis for V in this case, then using the row operations in this diagonalization as in the previous example we see that the generators of V corresponding to the factors above are $(x-1)e_1 - 2e_2 - e_3 = 0$, $-2e_1 + (x+1)e_2 - e_4 = 0$, e_1, e_2. Hence a vector space basis for the two direct factors in the invariant decomposition of V in this case is given by e_1, Te_1 and e_2, Te_2 where T is the linear transformation defined by D, i.e. $e_1, e_1 + 2e_2 + e_3$ and $e_2, 2e_1 - e_2 + e_4$. The corresponding matrix P relating these bases is

$$P = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

so that $P^{-1}DP$ is in rational canonical form:

$$P^{-1}DP = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

as can easily be checked.

II. (*Converting D directly to Rational Canonical Form*) As in Example 2 we determine the matrix P' of the algorithm from the row operations used in the diagonalization

of $xI - D$:

$$\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix} \xrightarrow[C_1 \leftrightarrow C_3]{} \begin{pmatrix}0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&0&1\end{pmatrix} \xrightarrow[-C_1]{} \begin{pmatrix}0&0&1&0\\0&1&0&0\\-1&0&0&0\\0&0&0&1\end{pmatrix} \rightarrow$$

$$\xrightarrow[\mapsto C_1]{C_1 - 2C_2} \begin{pmatrix}0&0&1&0\\-2&1&0&0\\-1&0&0&0\\0&0&0&1\end{pmatrix} \xrightarrow[\mapsto C_1]{C_1 + (D-I)C_3} \begin{pmatrix}0&0&1&0\\0&1&0&0\\0&0&0&0\\0&0&0&1\end{pmatrix} \xrightarrow[C_2 \leftrightarrow C_4]{} \begin{pmatrix}0&0&1&0\\0&0&0&1\\0&0&0&0\\0&1&0&0\end{pmatrix}$$

$$\xrightarrow[-C_2]{} \begin{pmatrix}0&0&1&0\\0&0&0&1\\0&0&0&0\\0&-1&0&0\end{pmatrix} \xrightarrow[\mapsto C_2]{C_2 - 2C_3} \begin{pmatrix}0&-2&1&0\\0&0&0&1\\0&0&0&0\\0&-1&0&0\end{pmatrix} \xrightarrow[\mapsto C_2]{C_2 + (D+I)C_4} \begin{pmatrix}0&0&1&0\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{pmatrix}$$

Here we have $d_1 = 2$ and $d_2 = 2$, corresponding to the third and fourth nonzero columns of P'. The columns of P are therefore given by

$$\begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad D\begin{pmatrix}1\\0\\0\\0\end{pmatrix} = \begin{pmatrix}1\\2\\1\\0\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad D\begin{pmatrix}0\\1\\0\\0\end{pmatrix} = \begin{pmatrix}2\\-1\\0\\1\end{pmatrix},$$

respectively, which again gives the matrix P above.

(4) In this example we determine all similarity classes of matrices A with entries from $\mathbb{Q}$ with characteristic polynomial $(x^4 - 1)(x^2 - 1)$. First note that any matrix with a degree 6 characteristic polynomial must be a 6×6 matrix. The polynomial $(x^4 - 1)(x^2 - 1)$ factors into irreducibles in $\mathbb{Q}[x]$ as $(x-1)^2(x+1)^2(x^2+1)$. Since the minimal polynomial $m_A(x)$ for A has the same roots as $c_A(x)$ it follows that $(x-1)(x+1)(x^2+1)$ divides $m_A(x)$. Suppose $a_1(x), \ldots, a_m(x)$ are the invariant factors of some A, so $a_m(x) = m_A(x)$, $a_i(x) \mid a_{i+1}(x)$ (in particular, all the invariant factors divide $m_A(x)$) and $a_1(x)a_2(x)\cdots a_m(x) = (x^4-1)(x^2-1)$. One easily sees that the only permissible lists under these constraints are

(a) $(x-1)(x+1), \quad (x-1)(x+1)(x^2+1)$
(b) $x-1, \quad (x-1)(x+1)^2(x^2+1)$
(c) $x+1, \quad (x-1)^2(x+1)(x^2+1)$
(d) $(x-1)^2(x+1)^2(x^2+1)$.

One can now easily write out the corresponding direct sums of companion matrices to obtain representatives of the 4 similarity classes. Note that there are still only 4 similarity classes even in $M_6(\mathbb{C})$ (cf. Exercise 13).

(5) In this example we find all similarity classes of 3×3 matrices A with entries from $\mathbb{Q}$ satisfying $A^6 = I$. For each such A, its minimal polynomial divides $x^6 - 1$ and in $\mathbb{Q}[x]$ the complete factorization of this polynomial is

$$x^6 - 1 = (x-1)(x+1)(x^2 - x + 1)(x^2 + x + 1).$$

Conversely, if B is any 3×3 matrix whose minimal polynomial divides $x^6 - 1$, then $B^6 = I$. The only restriction on the minimal polynomial for B is that its degree is at most 3 (by the Cayley–Hamilton Theorem). The only possibilities for the minimal

polynomial of such a matrix A are therefore

(a) $x - 1$
(b) $x + 1$
(c) $x^2 - x + 1$
(d) $x^2 + x + 1$
(e) $(x - 1)(x + 1)$
(f) $(x - 1)(x^2 - x + 1)$
(g) $(x - 1)(x^2 + x + 1)$
(h) $(x + 1)(x^2 - x + 1)$
(i) $(x + 1)(x^2 + x + 1)$.

Under the constraints of the rational canonical form these give rise to the following permissible lists of invariant factors:

(i) $x - 1, \quad x - 1, \quad x - 1$
(ii) $x + 1, \quad x + 1, \quad x + 1$
(iii) $x - 1, \quad (x - 1)(x + 1)$
(iv) $x + 1, \quad (x - 1)(x + 1)$
(v) $(x - 1)(x^2 - x + 1)$
(vi) $(x - 1)(x^2 + x + 1)$
(vii) $(x + 1)(x^2 - x + 1)$
(viii) $(x + 1)(x^2 + x + 1)$.

Note that it is impossible to have a suitable set of invariant factors if the minimal polynomial is $x^2 + x + 1$ or $x^2 - x + 1$. One can now write out the corresponding rational canonical forms; for example, (i) is I, (ii) is $-I$, and (iii) is

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

Note also that another way of phrasing this result is that any 3×3 matrix with entries from $\mathbb{Q}$ whose order (multiplicatively, of course) divides 6 is similar to one of these 8 matrices, so this example determines all elements of orders 1,2,3 and 6 in the group $\mathrm{GL}_3(\mathbb{Q})$ (up to similarity).

EXERCISES

1. Prove that similar linear transformations of V (or $n \times n$ matrices) have the same characteristic and the same minimal polynomial.
2. Let M be as in Lemma 19. Prove that the minimal polynomial of M is the least common multiple of the minimal polynomials of $A_1, \ldots, A_k$.
3. Prove that two 2×2 matrices over F which are not scalar matrices are similar if and only if they have the same characteristic polynomial.
4. Prove that two 3×3 matrices are similar if and only if they have the same characteristic and same minimal polynomials. Give an explicit counterexample to this assertion for 4×4 matrices.
5. Prove directly from the fact that the collection of *all* linear transformations of an n dimensional vector space V over F to itself form a vector space over F of dimension n^2 that the minimal polynomial of a linear transformation T has degree at most n^2.
6. Prove that the constant term in the characteristic polynomial of the $n \times n$ matrix A is $(-1)^n \det A$ and that the coefficient of x^{n-1} is the negative of the sum of the diagonal

entries of A (the sum of the diagonal entries of A is called the *trace* of A). Prove that $\det A$ is the product of the eigenvalues of A and that the trace of A is the sum of the eigenvalues of A.

7. Determine the eigenvalues of the matrix

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

8. Verify that the characteristic polynomial of the companion matrix

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & 0 & \dots & 0 & -a_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -a_{n-1} \end{pmatrix}$$

is

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

9. Find the rational canonical forms of

$$\begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} c & 0 & -1 \\ 0 & c & 1 \\ -1 & 1 & c \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 422 & 465 & 15 & -30 \\ -420 & -463 & -15 & 30 \\ 840 & 930 & 32 & -60 \\ -140 & -155 & -5 & 12 \end{pmatrix}.$$

10. Find all similarity classes of 6×6 matrices over $\mathbb{Q}$ with minimal polynomial $(x+2)^2(x-1)$ (it suffices to give all lists of invariant factors and write out some of their corresponding matrices).
11. Find all similarity classes of 6×6 matrices over $\mathbb{C}$ with characteristic polynomial $(x^4 - 1)(x^2 - 1)$.
12. Find all similarity classes of 3×3 matrices A over $\mathbb{F}_2$ satisfying $A^6 = I$ (compare with the answer we computed over $\mathbb{Q}$). Do the same for 4×4 matrices B satisfying $B^{20} = I$.
13. Prove that if there are k similarity classes of matrices over $\mathbb{Q}$ with a given characteristic polynomial in $\mathbb{Q}[x]$, then there are exactly the same number over any extension field of $\mathbb{Q}$. Is the same result true if "characteristic polynomial" is replaced by "minimal polynomial"?
14. Determine all possible rational canonical forms for a linear transformation with characteristic polynomial $x^2(x^2+1)^2$.
15. Determine up to similarity all 2×2 rational matrices (i.e. $\in M_2(\mathbb{Q})$) of precise order 4 (multiplicatively, of course). Do the same if the matrix has entries from $\mathbb{C}$.
16. Show that $x^5 - 1 = (x-1)(x^2 - 4x + 1)(x^2 + 5x + 1)$ in $\mathbb{F}_{19}[x]$. Use this to determine up to similarity all 2×2 matrices with entries from $\mathbb{F}_{19}$ of (multiplicative) order 5.
17. Determine representatives for the conjugacy classes for $GL_3(\mathbb{F}_2)$. [Compare your answer with Theorem 15 and Proposition 14 of Chapter 6.]
18. Let V be a finite dimensional vector space over $\mathbb{Q}$ and suppose T is a nonsingular linear transformation of V such that $T^{-1} = T^2 + T$. Prove that the dimension of V is divisible by 3. If the dimension of V is precisely 3 prove that all such transformations T are similar.

19. Let V be the infinite dimensional real vector space

$$\mathbb{R}^\infty = \{(a_0, a_1, a_2, \ldots) \mid a_0, a_1, a_2, \cdots \in \mathbb{R}\}.$$

Define the map $T : V \to V$ by $T(a_0, a_1, a_2, \ldots) = (0, a_0, a_1, a_2, \ldots)$. Prove that T has no eigenvectors.

20. Let ℓ be a prime and let $\Phi_\ell(x) = \frac{x^\ell - 1}{x-1} = x^{\ell-1} + x^{\ell-2} + \ldots + x + 1 \in \mathbb{Z}[x]$ be the ℓth cyclotomic polynomial, which is irreducible over $\mathbb{Q}$ (Example 4 following Corollary 9.15). This exercise determines the smallest degree of a factor of $\Phi_\ell(x)$ modulo p for any prime p and so in particular determines when $\Phi_\ell(x)$ is irreducible modulo p. (This actually determines the complete factorization of $\Phi_\ell(x)$ modulo p — cf. Exercise 8 of Section 13.6.)

(a) Show that if $p = l$ then $\Phi_\ell(x)$ is divisible by $x - 1$ in $\mathbb{F}_\ell[x]$.

(b) Suppose $p \neq \ell$ and let f denote the order of p in $\mathbb{F}_\ell^\times$, i.e. f is the smallest power of p with $p^f \equiv 1 \bmod \ell$. Show that $m = f$ is the first value of m for which the group $GL_m(\mathbb{F}_p)$ contains an element A of order ℓ. [Use the formula for the order of this group at the end of Section 11.1.]

(c) Show that $\Phi_\ell(x)$ is not divisible by any polynomial of degree smaller than f in $\mathbb{F}_p[x]$ [consider the companion matrix for such a divisor and use (b)]. Let $m_A(x) \in \mathbb{F}_p[x]$ denote the minimal polynomial for the matrix A in (b) and conclude that $m_A(x)$ is irreducible of degree f and divides $\Phi_\ell(x)$ in $\mathbb{F}_p[x]$.

(d) In particular, prove that $\Phi_\ell(x)$ is irreducible modulo p if and only if $l - 1$ is the smallest power of p which is congruent to 1 modulo ℓ, i.e. p is a primitive root modulo ℓ.

21. Prove that the first two elementary row and column operations described before Theorem 21 do not change the determinant of the matrix and the third elementary operation multiplies the determinant by a unit. Conclude from Theorem 21 that the characteristic polynomial of A differs by a unit from the product of the invariant factors of A. Since both these polynomials are monic by definition, conclude that they are equal (this gives an alternate proof of Proposition 20).

The following exercises outline the proof of Theorem 21. They carry out explicitly the construction described in the exercises at the end of the previous section for the Euclidean Domain $F[x]$. Let V be an n-dimensional vector space with basis $v_1, v_2, \ldots, v_n$ and let T be the linear transformation of V defined by the matrix A and this choice of basis, i.e. T is the linear transformation with

$$T(v_j) = \sum_{i=1}^{n} a_{ij} v_i \quad , \quad j = 1, 2, \ldots, n$$

where $A = (a_{ij})$. Let $F[x]^n$ be the free module of rank n over $F[x]$ and let $\xi_1, \xi_2, \ldots, \xi_n$ denote a basis. Then we have a natural surjective $F[x]$-module homomorphism

$$\varphi : F[x]^n \to V$$

defined by mapping ξ_i to v_i, $i = 1, 2, \ldots, n$. As indicated in the exercises of the previous section the invariant factors for the $F[x]$-module V can be determined once we have determined a set of generators and the corresponding relations matrix for $\ker \varphi$. Since by definition x acts on V

by the linear transformation T, we have

$$x(v_j) = \sum_{i=1}^{n} a_{ij} v_i \quad , \quad j = 1, 2, \ldots, n$$

22. Show that the elements

$$\nu_j = -a_{1j}\xi_1 - \cdots - a_{j-1\,j}\xi_{j-1} + (x - a_{jj})\xi_j - a_{j+1\,j}\xi_{j+1} - \cdots - a_{nj}\xi_n$$

for $j = 1, 2, \ldots, n$ are elements of the kernel of φ.

23. **(a)** Show that $x\xi_j = \nu_j + f_j$ where $f_j \in F\xi_1 + \cdots + F\xi_n$ is an element in the F-vector space spanned by $\xi_1, \ldots, \xi_n$.

(b) Show that

$$F[x]\xi_1 + \cdots + F[x]\xi_n = (F[x]\nu_1 + \cdots + F[x]\nu_n) + (F\xi_1 + \cdots + F\xi_n).$$

24. Show that $\nu_1, \nu_2, \ldots, \nu_n$ generate the kernel of φ. [Use the previous result to show that any element of ker φ is the sum of an element in the module generated by $\nu_1, \nu_2, \ldots, \nu_n$ and an element of the form $b_1\xi_1 + \cdots + b_n\xi_n$ where the b_i are elements of F. Then show that such an element is in ker φ if and only if all the b_i are 0 since $v_1, \ldots, v_n$ are a basis for V over F.]

25. Show that the generators $\nu_1, \nu_2, \ldots, \nu_n$ of ker φ have corresponding relations matrix

$$\begin{pmatrix} x - a_{11} & -a_{21} & \cdots & -a_{n1} \\ -a_{12} & x - a_{22} & \cdots & -a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{1n} & -a_{2n} & \cdots & x - a_{nn} \end{pmatrix} = xI - A^t \quad ,$$

where A^t is the transpose of A. Conclude that Theorem 21 and the algorithm for determining the invariant factors of A follows by the exercises at the end of the previous section (note that the row and column operations necessary to diagonalize this relations matrix are the column and row operations necessary to diagonalize the matrix in Theorem 21, which explains why the invariant factor algorithm keeps track of the *row* operations used).

12.3 THE JORDAN CANONICAL FORM

We continue with the notation in the previous section: F is a field, $F[x]$ is the ring of polynomials in x with coefficients in F, V is a finite dimensional vector space over F of dimension n, T is a fixed linear transformation of V by which we make V into an $F[x]$-module, and A is an $n \times n$ matrix with coefficients in F. Recall that once a basis for V has been fixed any linear transformation T defines a matrix A and conversely any matrix A defines a linear transformation T.

In the previous section we used the invariant factor form of the Fundamental Theorem for finitely generated modules over the Principal Ideal Domain $F[x]$ to ob-

tain the rational canonical form for such a linear transformation T and the rational canonical form for such an $n \times n$ matrix A. In this section we use the elementary divisor form of the Fundamental Theorem to obtain the *Jordan canonical form*. We shall see that matrices in this canonical form are as close to being diagonal matrices as possible, so the matrices are simpler than in the rational canonical form (but we lose some of the "rationality" results).

The elementary divisors of a module are the prime power divisors of its invariant factors (this was Corollary 10). For the $F[x]$-module V the invariant factors were monic polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ of degree at least one (with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$), so the associated elementary divisors are the powers of the irreducible polynomial factors of these polynomials. These polynomials are only defined up to multiplication by a unit and, as in the case of the invariant factors, we can specify them uniquely by requiring that they be monic.

To obtain the simplest possible elementary divisors we shall assume that the polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ factor completely into linear factors, i.e. that the elementary divisors of V are powers $(x - \lambda)^k$ of linear polynomials. Since the product of the elementary divisors is the characteristic polynomial, this is equivalent to the assumption that the field F contains all the eigenvalues of the linear transformation T (equivalently, of the matrix A representing the linear transformation T).

Under this assumption on F, it follows immediately from Theorem 6 that V is the direct sum of finitely many cyclic $F[x]$-modules of the form $F[x]/(x - \lambda)^k$ where $\lambda \in F$ is one of the eigenvalues of T, corresponding to the elementary divisors of V.

We now choose a vector space basis for each of the direct summands corresponding to the elementary divisors of V for which the corresponding matrix for T is particularly simple. Recall that by definition of the $F[x]$-module structure the linear transformation T acting on V is the element x acting by multiplication on each of the direct summands $F[x]/(x - \lambda)^k$.

Consider the elements

$$(\bar{x} - \lambda)^{k-1},\ (\bar{x} - \lambda)^{k-2}, \ldots,\ \bar{x} - \lambda,\ 1,$$

in the quotient $F[x]/(x - \lambda)^k$. Expanding each of these polynomials in $\bar{x}$ we see the matrix relating these elements to the F-basis $\bar{x}^{k-1}, \bar{x}^{k-2}, \ldots, \bar{x}, 1$ of $F[x]/(x - \lambda)^k$ is upper triangular with 1s along the diagonal. Since this is an invertible matrix (having determinant 1), it follows that the elements above are an F-basis for $F[x]/(x - \lambda)^k$. With respect to this basis the linear transformation of multiplication by x acts in a particularly simple manner (note that $x = \lambda + (x - \lambda)$ and that $(\bar{x} - \lambda)^k = 0$ in the quotient):

$$x: \quad \begin{array}{lcl} (\bar{x} - \lambda)^{k-1} & \mapsto & \lambda \cdot (\bar{x} - \lambda)^{k-1} + (\bar{x} - \lambda)^k = \lambda \cdot (\bar{x} - \lambda)^{k-1} \\ (\bar{x} - \lambda)^{k-2} & \mapsto & \lambda \cdot (\bar{x} - \lambda)^{k-2} + (\bar{x} - \lambda)^{k-1} \\ & \vdots & \\ \bar{x} - \lambda & \mapsto & \lambda \cdot (\bar{x} - \lambda) + (\bar{x} - \lambda)^2 \\ 1 & \mapsto & \lambda \cdot 1 + (\bar{x} - \lambda). \end{array}$$

With respect to this basis, the matrix for multiplication by x is therefore

$$\begin{pmatrix} \lambda & 1 & & & \\ & \lambda & \ddots & & \\ & & \ddots & 1 & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

where the blank entries are all zero. Such matrices are given a name:

Definition. The $k \times k$ matrix with λ along the main diagonal and 1 along the first superdiagonal depicted above is called the *$k \times k$ elementary Jordan matrix with eigenvalue λ* or the *Jordan block of size k with eigenvalue λ*.

Applying this to each of the cyclic factors of V in its elementary divisor decomposition we obtain a vector space basis for V with respect to which the linear transformation T has as matrix the direct sum of the Jordan blocks corresponding to the elementary divisors of V, i.e. is block diagonal with Jordan blocks along the diagonal:

$$\begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_t \end{pmatrix}.$$

Notice that this matrix is uniquely determined up to permutation of the blocks along the diagonal by the elementary divisors of the $F[x]$-module V and conversely, by Theorem 9, the list of elementary divisors uniquely determines the module V up to $F[x]$-module isomorphism.

Definition

(1) A matrix is said to be in *Jordan canonical form* if it is a block diagonal matrix with Jordan blocks along the diagonal.

(2) A *Jordan canonical form* for a linear transformation T is a matrix representing T which is in Jordan canonical form.

We have proved that any linear transformation T has a Jordan canonical form. As in the case of the rational canonical form, it follows from the uniqueness of the elementary divisors that the Jordan canonical form is unique up to a permutation of the Jordan blocks along the diagonal (hence is called *the* Jordan canonical form for T). We summarize this as:

Theorem 22. (Jordan Canonical Form for Linear Transformations) Let V be a finite dimensional vector space over the field F and let T be a linear transformation of V. Assume F contains all the eigenvalues of T.

(1) There is a basis for V with respect to which the matrix for T is in Jordan canonical form, i.e. is a block diagonal matrix whose diagonal blocks are the Jordan blocks for the elementary divisors of V.

(2) The Jordan canonical form for T is unique up to a permutation of the Jordan blocks along the diagonal.

As for the rational canonical form, there is a corresponding statement for $n \times n$ matrices over F:

Theorem 23. (Jordan Canonical Form for Matrices) Let A be an $n \times n$ matrix over the field F and assume F contains all the eigenvalues of A.

(1) The matrix A is similar to a matrix in Jordan canonical form, i.e. there is an invertible $n \times n$ matrix P over F such that $P^{-1}AP$ is a block diagonal matrix whose diagonal blocks are the Jordan blocks for the elementary divisors of A.

(2) The Jordan canonical form for A is unique up to a permutation of the Jordan blocks along the diagonal.

The Jordan canonical form differs from a diagonal matrix only by the possible presence of some 1's along the first superdiagonal (and then only if there are Jordan blocks of size greater than one), hence is close to being a diagonal matrix. The following result shows in particular that the Jordan canonical form for a matrix A is as close to being a diagonal matrix as possible.

Corollary 24

(1) If a matrix A is similar to a diagonal matrix D, then D is the Jordan canonical form of A.

(2) Two diagonal matrices are similar if and only if their diagonal entries are the same up to a permutation.

Proof: The first assertion is immediate from the uniqueness of Jordan canonical forms because a diagonal matrix is itself in Jordan form (with Jordan blocks of size 1). The uniqueness of the Jordan canonical form gives (2).

The next corollary gives a criterion to determine when a matrix A can be diagonalized.

Corollary 25. If A is an $n \times n$ matrix with entries from F and F contains all the eigenvalues of A, then A is similar to a diagonal matrix over F if and only if the minimal polynomial of A has no repeated roots.

Proof: Suppose A is similar to a diagonal matrix. The minimal polynomial of a diagonal matrix has no repeated roots (its roots are precisely the distinct elements along the diagonal). Since similar matrices have the same minimal polynomial it follows that the minimal polynomial for A has no repeated roots.

Conversely, suppose the minimal polynomial for A has no repeated roots and let B be the Jordan canonical form of A. The matrix B is a block diagonal matrix with elementary Jordan matrices down the diagonal. By the exercises at the end of the preceding section the minimal polynomial for B is the least common multiple of the minimal polynomials of the Jordan blocks. It is easy to see directly that a Jordan

block of size k with eigenvalue λ has minimal polynomial $(x-\lambda)^k$ (note that this is immediate from the fact that each elementary Jordan matrix gives the action on a *cyclic* $F[x]$-submodule whose annihilator is $(x-\lambda)^k$). Since A and B have the same minimal polynomial, the least common multiple of the $(x-\lambda)^k$ cannot have any repeated roots. It follows that k must be 1, i.e. that each Jordan block must be of size one and B is a diagonal matrix.

Changing From One Canonical Form to Another

We continue to assume that the field F contains all the eigenvalues of T (or A) so both the rational and Jordan canonical forms exist over F. The process of passing from one form to the other is exactly the same algorithm described in Section 5.2 for finite abelian groups (where the elementary divisors were determined from the list of invariant factors and vice versa).

In brief summary, recall that the elementary divisors are the prime power divisors of the invariant factors. They are obtained from the invariant factors by writing each invariant factor as a product of distinct linear factors to powers; the resulting set of powers of linear polynomials is the set of elementary divisors. For example, if the invariant factors of T are:

$$(x-1)(x-3)^3, \quad (x-1)(x-2)(x-3)^3, \quad (x-1)(x-2)^2(x-3)^3$$

then the elementary divisors are:

$$(x-1), \quad (x-3)^3, \quad (x-1), \quad (x-2), \quad (x-3)^3, \quad (x-1), \quad (x-2)^2, \quad (x-3)^3.$$

The largest invariant factor is the product of the largest of the distinct prime powers among the elementary divisors, the next largest invariant factor is the product of the largest of the distinct prime powers among the remaining elementary divisors, and so on. Given a list of elementary divisors we can find the list of invariant factors by first arranging the elementary divisors into n separate lists, one for each eigenvalue. In each of these n lists arrange the polynomials in increasing (i.e. nondecreasing) degree. Next arrange for all n lists to have the same length by appending an appropriate number of the constant polynomial 1. Now form the ith invariant factor by taking the product of the ith polynomial in each of these lists. For example, if the elementary divisors of T are:

$$(x-1)^3, (x+4), (x+4)^2, (x-5)^2, (x-1)^5, (x-1)^3, (x-5)^3, (x-1)^4, (x+4)^3$$

then the intermediate lists are

$$\begin{array}{llll} (1)\ (x-1)^3, & (x-1)^3, & (x-1)^4, & (x-1)^5 \\ (2)\ 1, & x+4, & (x+4)^2, & (x+4)^3 \\ (3)\ 1, & 1, & (x-5)^2, & (x-5)^3 \end{array}$$

so the list of invariant factors is:

$$(x-1)^3, \quad (x-1)^3(x+4), \quad (x-1)^4(x+4)^2(x-5)^2, \quad (x-1)^5(x+4)^3(x-5)^3.$$

Elementary Divisor Decomposition Algorithm: Converting to Jordan Canonical Forms

Theorem 21 indicates a computational procedure to determine the invariant factors of any given matrix A. Factorization of these invariant factors produces the elementary divisors of A, hence determines the Jordan canonical form for A as above.

The Invariant Factor Decomposition Algorithm following Theorem 21 starts with a basis $e_1, \ldots, e_n$ for V and produces a set $f_1, \ldots, f_m$ of elements of V which are $F[x]$-module generators for the cyclic factors in the invariant factor decomposition of V (with annihilators $(a_1(x)), \ldots, (a_m(x))$, respectively). Since the elementary divisor decomposition is obtained from the invariant factor decomposition by applying the Chinese Remainder Theorem to the cyclic modules $F[x]/(a_i(x))$, this gives a set of $F[x]$-module generators for the cyclic factors in the elementary divisor decomposition of V. These elements then give rise to an explicit vector space basis for V with respect to which the linear transformation corresponding to A is in Jordan canonical form (equivalently, an explicit matrix P such that $P^{-1}AP$ is in Jordan canonical form). As for the Invariant Factor Decomposition Algorithm we state the result first in the general context of decomposing a vector space and then describe the algorithm to convert a given $n \times n$ matrix A to Jordan canonical form.

Explicit numerical examples of this algorithm are given later in Examples 2 and 3.

Elementary Divisor Decomposition Algorithm

(1) to **(3)**: The first three steps in the algorithm are those from the Invariant Factor Decomposition Algorithm following Theorem 21.

(4) For each invariant factor $a(x)$ computed for A write

$$a(x) = (x - \lambda_1)^{\alpha_1}(x - \lambda_2)^{\alpha_2} \ldots (x - \lambda_s)^{\alpha_s}$$

where $\lambda_1, \ldots, \lambda_s \in F$ are distinct. Let $f \in V$ be the $F[x]$-module generator for the cyclic factor corresponding to the invariant factor $a(x)$ computed in (3). Then the elements

$$\frac{a(x)}{(x-\lambda_1)^{\alpha_1}} f, \quad \frac{a(x)}{(x-\lambda_2)^{\alpha_2}} f, \quad \ldots, \quad \frac{a(x)}{(x-\lambda_s)^{\alpha_s}} f$$

(note that the $\dfrac{a(x)}{(x-\lambda_i)^{\alpha_i}} \in F[x]$ are polynomials) are $F[x]$-module generators for the cyclic factors of V corresponding to the elementary divisors

$$(x-\lambda_1)^{\alpha_1}, \quad (x-\lambda_2)^{\alpha_2}, \quad \ldots, \quad (x-\lambda_s)^{\alpha_s},$$

respectively.

(5) If $g_i = \dfrac{a(x)}{(x-\lambda_i)^{\alpha_i}} f$ is the $F[x]$-module generator for the cyclic factor of V corresponding to the elementary divisor $(x-\lambda_i)^{\alpha_i}$ then the corresponding *vector*

space basis for this cyclic factor of V is given by the elements

$$(T-\lambda_i)^{\alpha_i-1}g_i, \quad (T-\lambda_i)^{\alpha_i-2}g_i, \quad \ldots, \quad (T-\lambda_i)g_i, \quad g_i.$$

(6) Write the kth element of the vector space basis computed in (5) in terms of the original vector space basis $[e_1, e_2, \ldots, e_n]$ for V and use the coordinates for the kth column of an $n \times n$ matrix P. Then $P^{-1}AP$ is in Jordan canonical form (with Jordan blocks appearing in the order used in (5) for the cyclic factors of V).

Converting an *n* x *n* Matrix to Jordan Canonical Form

(1) to **(2)**: The first two steps are those from the algorithm for Converting an $n \times n$ matrix to Rational Canonical Form following Theorem 21.

(3) When $xI - A$ has been diagonalized to the form in Theorem 21 the first $n-m$ columns of the matrix P' are 0 (providing a useful numerical check on the computations) and the remaining m columns of P' are nonzero. For each successive $i = 1, 2, \ldots, m$:

(a) Factor the ith nonconstant diagonal element (which is of degree d_i):

$$a(x) = (x-\lambda_1)^{\alpha_1}(x-\lambda_2)^{\alpha_2}\ldots(x-\lambda_s)^{\alpha_s}$$

where $\lambda_1, \ldots, \lambda_s \in F$ are distinct (here $a(x) = a_i(x)$ is the ith nonconstant diagonal element and s depends on i).

(b) Multiply the ith nonzero column of P' successively by the d_i matrices:

$$\begin{array}{lll}
(A-\lambda_1 I)^{\alpha_1-1} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{\alpha_s} \\
(A-\lambda_1 I)^{\alpha_1-2} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & \vdots & \\
(A-\lambda_1 I)^{0} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & & \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2-1} & \ldots(A-\lambda_s I)^{\alpha_s} \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2-2} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & \vdots & \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{0} & \ldots(A-\lambda_s I)^{\alpha_s} \\
 & & \\
 & \vdots & \\
 & & \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{\alpha_s-1} \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{\alpha_s-2} \\
 & \vdots & \\
(A-\lambda_1 I)^{\alpha_1} & (A-\lambda_2 I)^{\alpha_2} & \ldots(A-\lambda_s I)^{0}.
\end{array}$$

(c) Use the column vectors resulting from (b) (in that order) as the next d_i columns of an $n \times n$ matrix P.

Then $P^{-1}AP$ is in Jordan canonical form (whose Jordan blocks correspond to the ordering of the factors in (a)).

Examples

We can use Jordan canonical forms to carry out the same analysis of matrices that we did as examples of the use of rational canonical forms. In some instances, when the field is enlarged, the number of similarity classes increases (the number of similarity classes can never decrease when we extend the field by Corollary 18(2)).

(1) Let A, B and C be the matrices in Example 1 of the previous section and let $F = \mathbb{Q}$. Note that $\mathbb{Q}$ contains all the eigenvalues for these matrices. Since we have already determined the invariant factors of these matrices we can immediately obtain their elementary divisors. The elementary divisors of A are $x-2$, $x-2$ and $x-3$ and the elementary divisors of B and C are $(x-2)^2$ and $x-3$ so the respective Jordan canonical forms are:

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \qquad \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Notice that A is similar to a diagonal matrix but, by Corollary 25, B and C are not.

(2) For the matrix A, we determined in Example 2 of the previous section that $f_1 = -7e_1 + 7e_2 + e_3$ and $f_2 = -e_1 + e_2$ were $\mathbb{Q}[x]$-module generators for the two cyclic factors of V in its invariant factor decomposition, corresponding to the invariant factors $x-2$ and $(x-2)(x-3)$, respectively. Using the first algorithm described above, the elements f_1, $(x-3)f_2$ and $(x-2)f_2$ are therefore $\mathbb{Q}[x]$-module generators for the three cyclic factors of V in its elementary divisor decomposition, corresponding to the elementary divisors $x-2$, $x-2$, and $x-3$. An easy computation shows that these are the elements $-7e_1 + 7e_2 + e_3$, $-e_1$ and $-2e_1 + e_2$, respectively. Then the matrix

$$P = \begin{pmatrix} -7 & -1 & -2 \\ 7 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

conjugates A into its Jordan canonical form:

$$P^{-1}AP = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

as one easily checks.

The columns of this matrix can also be obtained following the second algorithm above, using the nonzero columns of the matrix P' computed in Example 2 of the previous section:

$$(A-2I)^0 \begin{pmatrix} -7 \\ 7 \\ 1 \end{pmatrix} = \begin{pmatrix} -7 \\ 7 \\ 1 \end{pmatrix}$$

and

$$(A-2I)^0(A-3I)^1 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \quad (A-2I)^1(A-3I)^0 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix},$$

respectively, which again gives the matrix P.

(3) For the 4×4 matrix D of Example 3 of the previous section, the invariant factors were $(x-1)^2$, $(x-1)^2$, with corresponding $\mathbb{Q}[x]$-module generators $f_1 = e_1$ and $f_2 = e_2$, respectively. These are also the elementary divisors for this matrix. The corresponding vector space bases for these two factors are given by $(T-1)f_1, f_1$ and $(T-1)f_2, f_2$, respectively. An easy computation shows these are the elements $2e_2 + e_3, e_1$ and $2e_1 - e_2 + e_4, e_2$, respectively. Then the matrix

$$P = \begin{pmatrix} 0 & 1 & 2 & 0 \\ 2 & 0 & -2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

conjugates D into its Jordan canonical form:

$$P^{-1}DP = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

as can easily be checked.

The columns of this matrix can also be obtained following the second algorithm above, using the nonzero columns of the matrix P' computed in Example 3 of the previous section:

$$(D-I)^1 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 1 \\ 0 \end{pmatrix}, \qquad (D-I)^0 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and

$$(D-I)^1 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ 0 \\ 1 \end{pmatrix}, \qquad (D-I)^0 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix},$$

respectively, which again gives the matrix P.

(4) The set of similarity classes of 6×6 matrices with entries from $\mathbb{C}$ with characteristic polynomial $(x^4-1)(x^2-1)$ consists of the 4 classes represented by the rational canonical forms in the preceding set of examples (there are no additional lists of invariant factors over $\mathbb{C}$). Their Jordan canonical forms cannot all be written over $\mathbb{Q}$, however. For instance, if the invariant factors are

$$(x-1)(x+1) \quad \text{and} \quad (x-1)(x+1)(x^2+1)$$

then the elementary divisors are

$$x-1, \quad x+1, \quad x-1, \quad x+1, \quad x-i, \quad x+i,$$

where i is a square root of -1 in $\mathbb{C}$, so the Jordan form for this matrix is a diagonal matrix with diagonal entries $1, 1, -1, -1, i, -i$.

(5) In contrast, the set of similarity classes of 3×3 matrices, A, over $\mathbb{C}$ satisfying $A^6 = I$ is considerably larger than that over $\mathbb{Q}$. If A is any such matrix, $m_A(x) \mid x^6 - 1$ so since the latter polynomial has no repeated roots in $\mathbb{C}$, the minimal polynomial of A has no repeated roots. By Corollary 25 the Jordan canonical form of A is a di-

agonal matrix. Since this diagonal matrix has the same minimal polynomial, its 6th power is also the identity, and so each diagonal entry is a 6th root of unity. For each list $\zeta_1, \zeta_2, \zeta_3$ of 6th roots of unity we obtain a Jordan canonical form, and two such forms are the same (i.e. give rise to similar matrices) if and only if the lists are permuted versions of each other. One finds that there are, up to similarity, 56 classes of such A's.

EXERCISES

1. Suppose the vector space V is the direct sum of cylic $F[x]$-modules whose annihilators are $(x+1)^2$, $(x-1)(x^2+1)^2$, (x^4-1) and $(x+1)(x^2-1)$. Determine the invariant factors and elementary divisors for V.
2. Prove that if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the $n \times n$ matrix A then ${\lambda_1}^k, \ldots, {\lambda_n}^k$ are the eigenvalues of A^k for any $k \geq 0$.
3. Use the method of Example 2 above to determine explicit matrices P_1 and P_2 with $P_1^{-1}BP_1$ and $P_2^{-1}CP_2$ in Jordan canonical form. Use this to explicitly construct a matrix Q which conjugates B into C (proving directly that these matrices are similar).
4. Prove that the Jordan canonical form for the matrix

$$\begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}$$

is that stated at the beginning of this chapter. Explicitly determine a matrix P which conjugates this matrix to its Jordan canonical form. Explain why this matrix cannot be diagonalized.

5. Compute the Jordan canonical form for the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & 1 & 3 \end{pmatrix}.$$

6. Determine which of the following matrices are similar:

$$\begin{pmatrix} -1 & 4 & -4 \\ 2 & -1 & 3 \\ 0 & -4 & 3 \end{pmatrix} \quad \begin{pmatrix} -3 & -4 & 0 \\ 2 & 3 & 0 \\ 8 & 8 & 1 \end{pmatrix} \quad \begin{pmatrix} 3 & 2 & -4 \\ 2 & 1 & 0 \\ 3 & -1 & 3 \end{pmatrix} \quad \begin{pmatrix} -1 & 4 & -4 \\ 0 & -3 & 2 \\ 0 & -4 & 3 \end{pmatrix}.$$

7. Determine the Jordan canonical forms for the following matrices:

$$\begin{pmatrix} 5 & 4 & 1 \\ -1 & 0 & 0 \\ -3 & -4 & 1 \end{pmatrix} \quad \begin{pmatrix} 3 & 4 & 2 \\ -2 & -3 & -1 \\ -4 & -4 & -3 \end{pmatrix}.$$

8. Prove that the matrices

$$A = \begin{pmatrix} 5 & 6 & 0 \\ -3 & -4 & 0 \\ -2 & 0 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 3 & -1 & 2 \\ -10 & 6 & -14 \\ -6 & 3 & -7 \end{pmatrix}$$

are similar. Prove that both A and B can be diagonalized and determine explicit matrices P_1 and P_2 with $P_1^{-1}AP_1$ and $P_2^{-1}BP_2$ in diagonal form.

9. Prove that the matrices

$$\begin{pmatrix} -8 & -10 & -1 \\ 7 & 9 & 1 \\ 3 & 2 & 0 \end{pmatrix} \quad B = \begin{pmatrix} -3 & 2 & -4 \\ 4 & -1 & 4 \\ 4 & -2 & 5 \end{pmatrix}$$

both have $(x-1)^2(x+1)$ as characteristic polynomial but that one can be diagonalized and the other cannot. Determine the Jordan canonical form for both matrices.

10. Find all Jordan canonical forms of 2×2, 3×3 and 4×4 matrices over $\mathbb{C}$.

11. Verify that the characteristic polynomial of

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & -2 & 0 & 1 \\ -2 & 0 & -1 & -2 \end{pmatrix}$$

is a product of linear factors over $\mathbb{Q}$. Determine the rational and Jordan canonical forms for A over $\mathbb{Q}$.

12. Determine the Jordan canonical form for the matrix

$$\begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

13. Determine the Jordan canonical form for the matrix

$$\begin{pmatrix} 3 & 0 & -2 & -3 \\ 4 & -8 & 14 & -15 \\ 2 & -4 & 7 & -7 \\ 0 & 2 & -4 & 3 \end{pmatrix}.$$

14. Prove that the matrices

$$A = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -4 & -1 & -4 & 0 \\ 2 & 1 & 3 & 0 \\ -2 & 4 & 9 & 1 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 0 & -4 & -7 \\ 3 & -8 & 15 & -13 \\ 2 & -4 & 7 & -7 \\ 1 & 2 & -5 & 1 \end{pmatrix}$$

are similar.

15. Prove that the matrices

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 5 & 2 & -8 & -8 \\ -6 & -3 & 8 & 8 \\ -3 & -1 & 3 & 4 \\ 3 & 1 & -4 & -5 \end{pmatrix}$$

both have characteristic polynomial $(x-3)(x+1)^3$. Determine whether they are similar and determine the Jordan canonical form for each matrix.

16. Determine the Jordan canonical form for the matrix

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

and determine a matrix P which conjugates this matrix into its Jordan canonical form.

17. Prove that any matrix A is similar to its transpose A^t.

18. Determine all possible Jordan canonical forms for a linear transformation with characteristic polynomial $(x-2)^3(x-3)^2$.

19. Prove that all $n \times n$ matrices with characteristic polynomial $f(x)$ are similar if and only if $f(x)$ has distinct roots.

20. Show that the following matrices are similar in $M_p(\mathbb{F}_p)$ ($p \times p$ matrices with entries from $\mathbb{F}_p$):

$$\begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

21. Show that if $A^2 = A$ then A is similar to a diagonal matrix which has only 0's and 1's along the diagonal.

22. Prove that an $n \times n$ matrix A with entries from $\mathbb{C}$ satisfying $A^3 = A$ can be diagonalized. Is the same statement true over *any* field F?

23. Suppose A is a 2×2 matrix with entries from $\mathbb{Q}$ for which $A^3 = I$ but $A \neq I$. Write A in rational canonical form and in Jordan canonical form viewed as a matrix over $\mathbb{C}$.

24. Prove there are no 5×5 matrices A over $\mathbb{Q}$ with $A^8 = I$ but $A^4 \neq I$.

25. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{Q}$ whose entries are all equal to 1.

26. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{F}_p$ whose entries are all equal to 1 (the answer depends on whether or not p divides n).

27. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{Q}$ whose entries are all equal to 1 except that the entries along the main diagonal are all equal to 0.

28. Determine the Jordan canonical form for the $n \times n$ matrix over $\mathbb{F}_p$ whose entries are all equal to 1 except that the entries along the main diagonal are all equal to 0.

The direct sum of the cyclic submodules of V corresponding to all the elementary divisors of V which are powers of the same $x - \lambda$ is called the *generalized eigenspace of* T corresponding to the eigenvalue λ. Note that this is the p-primary component of V for the prime $p = x - \lambda$ of $F[x]$ and consists of the elements of V which are annihilated by some power of the linear transformation $T - \lambda$. The matrix for T on the generalized eigenspace for λ is the block diagonal matrix of all Jordan blocks for T with the same eigenvalue λ.

29. Give an explicit example of a linear transformation of V for which the generalized eigenspaces do not equal the eigenspaces.

30. Suppose V_i is the generalized eigenspace corresponding to eigenvalue λ_i. Prove that the dimension of the eigenspace of T on V_i equals the number of Jordan blocks corresponding to eigenvalue λ.

31. Let N be an $n \times n$ matrix with coefficients in the field F. The matrix N is said to be *nilpotent* if some power of N is the zero matrix, i.e. $N^k = 0$ for some k. Prove that any nilpotent matrix is similar to a block diagonal matrix whose blocks are matrices with 1's along the first superdiagonal and 0's elsewhere.

32. Prove that if N is an $n \times n$ nilpotent matrix then in fact $N^n = 0$.

33. Let A be a strictly upper triangular $n \times n$ matrix (all entries on and below the main diagonal are zero). Prove that A is nilpotent.

34. Prove that the trace of a nilpotent $n \times n$ matrix is 0 (recall the trace of a matrix is the sum of the diagonal elements).

35. For $0 \leq i \leq n$, led d_i be the g.c.d. of the determinants of all the $i \times i$ minors of $xI - A$ for A as in Theorem 21 (take the 0×0 minor to be 1). Prove that the ith element along the diagonal of the Smith Normal Form for A is d_i/d_{i-1}. This gives the invariant factors for A. [Show these g.c.d.'s do not change under elementary row and column operations.]

36. Let $V = \mathbb{C}^n$ be the usual n-dimensional vector space of n-tuples $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ of complex numbers. Let T be the linear transformation defined by $T(\alpha_1, \alpha_2, \ldots, \alpha_n) = (0, \alpha_1, \alpha_2, \ldots, \alpha_{n-1})$. Determine the Jordan canonical form for T.

37. Let J be a Jordan block of size n with eigenvalue λ over $\mathbb{C}$.
    (a) Prove that the Jordan canonical form for the matrix J^2 is the Jordan block of size n with eigenvalue λ^2 if $\lambda \neq 0$.
    (b) If $\lambda = 0$ prove that the Jordan canonical form for J^2 has two blocks (with eigenvalues 0) of size $\frac{n}{2}, \frac{n}{2}$ if n is even and of size $\frac{n-1}{2}, \frac{n+1}{2}$ if n is odd.

38. Determine necessary and sufficient conditions for a matrix $A \in M_n(\mathbb{C})$ to have a square root, i.e. for there to exist another matrix $B \in M_n(\mathbb{C})$ such that $A = B^2$. [Suppose B is in Jordan canonical form and consider the Jordan canonical form for B^2 using the previous exercise.]

39. Let J be a Jordan block of size n with eigenvalue λ over a field F of characteristic 2. Determine the Jordan canonical form for the matrix J^2. Determine necessary and sufficient conditions for a matrix $A \in M_n(F)$ to have a square root, i.e. for there to exist another matrix $B \in M_n(F)$ such that $A = B^2$.

The remaining exercises explore functions (power series) of a matrix and introduce some applications of the Jordan canonical form to the theory of differential equations.

Throughout these exercises the matrices are assumed to be $n \times n$ matrices with entries from the field K, where K is either the real or complex numbers. Let

$$G(x) = \sum_{k=0}^{\infty} \alpha_k x^k$$

be a power series with coefficients from K. Let $G_N(x) = \sum_{k=0}^{N} \alpha_k x^k$ be the Nth partial sum of $G(x)$ and for each $A \in M_n(K)$ let $G_N(A)$ be the element of $M_n(K)$ obtained (as usual) by substituting A in this polynomial. For each fixed i, j we obtain a sequence of real or complex numbers c_{ij}^N, $N = 0, 1, 2, \ldots$ by taking c_{ij}^N to be the i, j entry of the matrix $G_N(A)$. The series

$$G(A) = \sum_{k=0}^{\infty} \alpha_k A^k$$

is said to *converge* to the matrix C in $M_n(K)$ if for each $i, j \in \{1, 2, \ldots, n\}$ the sequence c_{ij}^N, $N = 0, 1, 2, \ldots$ converges to the i, j entry of C (in which case we write $G(A) = C$). Say $G(A)$ *converges* if there is some $C \in M_n(K)$ such that $G(A) = C$. If A is a 1×1 matrix, this is the usual notion of convergence of a series in K.

For $A = (a_{ij}) \in M_n(K)$ define

$$||A|| = \sum_{i,j=1}^{n} |a_{ij}|$$

i.e. $||A||$ is the sum of the absolute values of all the entries of A.

40. Prove that for all $A, B \in M_n(K)$ and all $\alpha \in K$
(a) $||A + B|| \leq ||A|| + ||B||$
(b) $||AB|| \leq ||A|| \cdot ||B||$
(c) $||\alpha A|| = |\alpha| \cdot ||A||$.

41. Let R be the radius of convergence of the real or complex power series $G(x)$ (where $R = \infty$ if $G(x)$ converges for all $x \in K$).
(a) Prove that if $||A|| < R$ then $G(A)$ converges.
(b) Deduce that for *all* matrices A the following power series converge:

$$\sin(A) = A - \frac{A^3}{3!} + \frac{A^5}{5!} + \cdots + (-1)^k \frac{A^{2k+1}}{(2k+1)!} + \cdots$$

$$\cos(A) = I - \frac{A^2}{2!} + \frac{A^4}{4!} + \cdots + (-1)^k \frac{A^{2k}}{(2k)!} + \cdots$$

$$\exp(A) = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots + \frac{A^k}{k!} + \cdots$$

where I is the $n \times n$ identity matrix.

In view of applications to the theory of differential equations we introduce a variable t at this point, so that for $A \in M_n(K)$ the matrix At is obtained from A by multiplying each entry by t (which is the same as multiplying A by the "scalar" matrix tI). We obtain a function from a subset of K into $M_n(K)$ defined by $t \mapsto G(At)$ at all points t where the series $G(At)$ converges. In particular, $\sin(At)$, $\cos(At)$ and $\exp(At)$ converge for all $t \in K$.

42. Let P be a nonsingular $n \times n$ matrix.
(a) Prove that $PG(At)P^{-1} = G(PAtP^{-1}) = G(PAP^{-1}t)$. (This implies that, up to a change of basis, it suffices to compute $G(At)$ for matrices A in canonical form). [Take limits of partial sums to get the first equality. The second equality is immediate because the matrix tI commutes with every matrix.]
(b) Prove that if A is the direct sum of matrices $A_1, A_2, \ldots, A_m$, then $G(At)$ is the direct sum of the matrices $G(A_1t), G(A_2t), \ldots, G(A_mt)$.
(c) Show that if Z is the diagonal matrix with entries $z_1, z_2, \ldots, z_n$ then $G(Zt)$ is the diagonal matrix with entries $G(z_1t), G(z_2t), \ldots, G(z_nt)$.

The matrix $\exp(A)$ defined in Exercise 41(b) is called the *exponential* of A and is often denoted by e^A. The next three exercises lead to a formula for the matrix $\exp(Jt)$, where J is an elementary Jordan matrix.

43. Prove that if A and B are *commuting* matrices then $\exp(A + B) = \exp(A)\exp(B)$. [Treat A and B as commuting indeterminates and deduce this by comparing the power series on the left hand side with the product of the two power series on the right hand side.]

44. Use the preceding exercise to show that if M is any matrix and λ is any element of K then

$$\exp(\lambda It + M) = e^{\lambda t}\exp(M).$$

45. Let N be the $r \times r$ matrix with 1's on the first superdiagonal and zeros elsewhere. Compute the exponential of the following nilpotent $r \times r$ matrix:

$$\text{if } Nt = \begin{pmatrix} 0 & t & & & \\ & 0 & t & & \\ & & \ddots & & \\ & & & & t \\ & & & & 0 \end{pmatrix} \quad \text{then } \exp(Nt) = \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \cdots & \frac{t^{r-1}}{(r-1)!} \\ & 1 & t & \frac{t^2}{2!} & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & t & \frac{t^2}{2!} \\ & & & & 1 & t \\ & & & & & 1 \end{pmatrix}.$$

Deduce that if J is the $r \times r$ elementary Jordan matrix with eigenvalue λ then

$$\exp(Jt) = \begin{pmatrix} e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & \cdots & \cdots & \frac{t^{r-1}}{(r-1)!}e^{\lambda t} \\ & e^{\lambda t} & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} & & \vdots \\ & & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & te^{\lambda t} & \frac{t^2}{2!}e^{\lambda t} \\ & & & & e^{\lambda t} & te^{\lambda t} \\ & & & & & e^{\lambda t} \end{pmatrix}.$$

[To do the first part use the observation that since Nt is a nilpotent matrix, $\exp(Nt)$ is a *polynomial* in Nt, i.e. all but a finite number of the terms in the power series are zero. To compute the exponential of Jt write Jt as $\lambda It + Nt$ and use Exercise 44 with $M = Nt$.]

Let $A \in M_n(K)$ and let P be a change of basis matrix such that $P^{-1}AP$ is in Jordan canonical form. Suppose $P^{-1}AP$ is the sum of elementary Jordan matrices $J_1, \ldots, J_m$. The preceding exercises (with $t = 1$) show that $\exp(A)$ can easily be found by writing $E = \exp(P^{-1}AP)$ as the direct sum of the matrices $\exp(J_1), \ldots, \exp(J_m)$ and then changing the basis back again to obtain $\exp(A) = PEP^{-1}$.

46. For the 4×4 matrices D and P given in Example 3 of this section:

$$D = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix} \qquad P = \begin{pmatrix} 0 & 1 & 2 & 0 \\ 2 & 0 & -2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

show that

$$E = \begin{pmatrix} e & e & 0 & 0 \\ 0 & e & 0 & 0 \\ 0 & 0 & e & e \\ 0 & 0 & 0 & e \end{pmatrix} \quad \text{and} \quad \exp(D) = \begin{pmatrix} e & 2e & -4e & 4e \\ 2e & -e & 4e & -8e \\ e & 0 & e & -2e \\ 0 & e & -2e & 3e \end{pmatrix}.$$

47. Compute the exponential of each of the following matrices:
(a) the matrix A in Example 2 of this section

(b) the matrix in Exercise 4 (where you computed the Jordan canonical form and a change of basis matrix)

(c) the matrix in Exercise 16.

48. Show that $\exp(0) = I$ (here 0 is the zero matrix and I is the identity matrix). Deduce that $\exp(A)$ is nonsingular with inverse $\exp(-A)$ for all matrices $A \in M_n(K)$.

49. Prove that $\det(\exp(A)) = e^{\operatorname{tr}(A)}$, where $\operatorname{tr}(A)$ is the trace of A (the sum of the diagonal entries of A).

50. Fix any $A \in M_n(K)$. Prove that the map

$$K \to GL_n(K) \qquad \text{defined by} \qquad t \mapsto \exp(At)$$

is a group homomorphism (here K is the additive group of the field). (Note how this generalizes the familiar exponential map from K to $K^\times$, which is the $n = 1$ case. The subgroup $\{\exp(At) \mid t \in K\}$ is called a *1-parameter subgroup* of $GL_n(K)$. These subgroups and the exponential map play an important role in the theory of *Lie groups* – $GL_n(K)$ being a particular example of a Lie group.).

Let $G(x)$ be a power series having an infinite radius of convergence and fix a matrix $A \in M_n(K)$. The entries of the matrix $G(At)$ are K-valued functions of the variable t that are defined for all t. Let $c_{ij}(t)$ be the function of t in the i, j entry of $G(At)$. The *derivative* of $G(At)$ with respect to t, denoted by $\frac{d}{dt}G(At)$, is the matrix whose i, j entry is $\frac{d}{dt}c_{ij}(t)$ obtained by differentiating each of the entries of $G(At)$. In other words, if we identify $M_n(K)$ with K^{n^2} by considering each $n \times n$ matrix as an n^2-tuple, then $t \mapsto G(At)$ is a map from K to K^{n^2} (i.e. is a vector valued function of t) whose derivative is just the usual (componentwise) derivative of this vector valued function.

51. Establish the following properties of derivatives.

(a) If $G(x) = \sum_{k=0}^{\infty} \alpha_k x^k$ then $\frac{d}{dt}G(At) = A \sum_{k=1}^{\infty} k\alpha_k (At)^{k-1}$.

(b) If v is an $n \times 1$ matrix with (constant) entries from K then

$$\frac{d}{dt}(G(At)v) = \left(\frac{d}{dt}G(At)\right) v.$$

52. Deduce from Part (a) of the preceding exercise that

$$\frac{d}{dt}\exp(At) = A\,\exp(At).$$

Now let $y_1(t), \ldots, y_n(t)$ be differentiable functions of the real variable t that are related by the following linear system of first order differential equations with constant coefficients $a_{ij} \in K$:

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n \\ y_2' &= a_{21}y_1 + a_{22}y_2 + \ldots + a_{2n}y_n \\ &\vdots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n \end{aligned} \qquad (*)$$

(here the primes denote derivatives with respect to t). Let A be the matrix whose i, j entry is a_{ij}, so that $(*)$ may be written as

$$\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} = A \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

or, more succinctly, as $y' = Ay$, where y is the column vector of functions $y_1(t), \ldots, y_n(t)$.

An $n \times n$ matrix whose entries are functions of t and whose columns are independent solutions to the system $(*)$ is called a *fundamental matrix* of $(*)$. By the theory of differential equations, the set of vectors y that are solutions to the system $(*)$ form an n-dimensional vector space over K and so the columns of a fundamental matrix are a *basis for the vector space of all solutions to* $(*)$.

53. Prove that $\exp(At)$ is a fundamental matrix of $(*)$. Show also that if C is the $n \times 1$ constant vector whose entries are $y_1(0), \ldots, y_n(0)$ then $y(t) = \exp(At)C$ is the particular solution to the system $(*)$ satisfying the initial condition $y(0) = C$. (Note how this generalizes the 1-dimensional result that the single differential equation $y' = ay$ has e^{at} as a basis for the 1-dimensional space of solutions and the unique solution to this differential equation satisfying the initial condition $y(0) = c$ is $y = ce^{at}$.) [Use the preceding exercises.]

54. Prove that if M is a fundamental matrix of $(*)$ and if Q is a nonsingular matrix in $M_n(K)$, then MQ is also a fundamental matrix of $(*)$. [The columns of MQ are linear combinations of the columns of M.]

Now apply the preceding two exercises to solve some specific systems of differential equations as follows: given the matrix A in a system $(*)$, calculate a change of basis matrix P such that $B = P^{-1}AP$ is in Jordan canonical form. Then $\exp(At) = P\exp(Bt)P^{-1}$ is a fundamental matrix for $(*)$. By the preceding exercise, $P\exp(Bt)$ is also a fundamental matrix for $(*)$ and $\exp(Bt)$ can be calculated by the method described in the discussion following Exercise 45 (in particular, one does not have to find the inverse of the matrix P to obtain a fundamental matrix for $(*)$). Thus, for example, if $A = D$ and P are the matrices given in Exercise 46, then we saw that the Jordan canonical form for A is the matrix $B = P^{-1}AP$ consisting of two 2×2 Jordan blocks with eigenvalues 1. A fundamental matrix for the system $y' = Ay$ is therefore

$$P\exp(B) = \begin{pmatrix} 0 & 1 & 2 & 0 \\ 2 & 0 & -2 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} e^t & te^t & 0 & 0 \\ 0 & e^t & 0 & 0 \\ 0 & 0 & e^t & te^t \\ 0 & 0 & 0 & e^t \end{pmatrix} = \begin{pmatrix} 0 & e^t & 2e^t & 2te^t \\ 2e^t & 2te^t & -2e^t & e^t(1-2t) \\ e^t & te^t & 0 & 0 \\ 0 & 0 & e^t & te^t \end{pmatrix}.$$

Writing this out more explicitly, this shows that the general solution to the system of differential equations:

$$\begin{aligned} y_1' &= y_1 + 2y_2 - 4y_3 + 4y_4 \\ y_2' &= 2y_1 - y_2 + 4y_3 - 8y_4 \\ y_3' &= y_1 + y_3 - 2y_4 \\ y_4' &= y_2 - 2y_3 + 3y_4 \end{aligned}$$

is given by:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \alpha_1 \begin{pmatrix} 0 \\ 2e^t \\ e^t \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} e^t \\ 2te^t \\ te^t \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 2e^t \\ -2e^t \\ 0 \\ e^t \end{pmatrix} + \alpha_4 \begin{pmatrix} 2te^t \\ e^t(1-2t) \\ 0 \\ te^t \end{pmatrix}$$

where $\alpha_1, \ldots, \alpha_4$ are arbitrary elements of the field K (this describes the 4-dimensional vector space of solutions).

55. In each of Parts (a) to (c) find a fundamental matrix for the system $(*)$, where the coefficient matrix A of $(*)$ is specified.
(a) A is the matrix in Part (a) of Exercise 47.
(b) A is the matrix in Part (b) of Exercise 47.
(c) A is the matrix in Part (c) of Exercise 47.

56. Consider the system $(*)$ whose coefficient matrix A is the matrix D listed in Exercise 46 and whose fundamental matrix was computed just before the preceding exercise. Find the particular solution to $(*)$ that satisfies the initial condition $y_i(0) = 1$ for $i = 1, 2, 3, 4$.

Next we explore a special case of $(*)$. Given the linear n^{th} *order* differential equation with constant coefficients

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0 \qquad (**)$$

(where $y^{(k)}$ is the kth derivative of y and $y^{(0)} = y$) one can form a *system* of linear *first order* differential equations by letting $y_i = y^{(i-1)}$ for $1 \le i \le n$ (the coefficient matrix of this system is described in the next exercise). A basis for the n-dimensional vector space of solutions to the nth order equation $(**)$ may then be obtained from a fundamental matrix for the linear system. Specifically, in each of the $n \times 1$ columns of functions in a fundamental matrix for the system, the $1,1$ entry is a solution to $(**)$ and so the n functions in the first row of the fundamental matrix for the system form a basis for the solutions to $(**)$.

57. Prove that the matrix, A, of coefficients of the system of n first order equations obtained from $(**)$ is the transpose of the companion matrix of the polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$.

58. Use the above methods to find a basis for the vector space of solutions to the following differential equations
(a) $y''' - 3y' + 2y = 0$
(b) $y'''' + 4y''' + 6y'' + 4y' + y = 0$.

A system of differential equations

$$\begin{aligned} y_1' &= F_1(y_1, y_2, \ldots, y_n) \\ y_2' &= F_2(y_1, y_2, \ldots, y_n) \\ &\vdots \\ y_n' &= F_n(y_1, y_2, \ldots, y_n) \end{aligned}$$

where $F_1, F_2, \ldots, F_n$ are functions of n variables, is called an *autonomous* system and it will be written more succinctly as $y' = F(y)$, where $F = (F_1, \ldots, F_n)$. (The expression autonomous means "independent of time" and it indicates that the variable t — which may be thought of as a time variable — does not appear explicitly on the right hand side.) The system $(*)$ is the special type of autonomous system in which each F_i is a linear function. In many instances it is desirable to analyze the behavior of solutions to an autonomous system of differential equations without explicitly finding these solutions (indeed, it is unlikely that it will be possible to find explicit solutions for a given nonlinear system). This investigation falls under the rubric "qualitative analysis" of autonomous differential equations and the rudiments of this study are often treated in basic calculus courses for 1×1 systems. The first step in a qualitative analysis of an $n \times n$ autonomous system is to find the *steady states*, namely the constant solutions (these are called steady states since they do not change with t). Note that a constant function $y = c$, where c is the $n \times 1$ constant vector with entries $c_1, \ldots, c_n$, is a solution to $y' = F(y)$ if and only if

$$c_i' = 0 = F_i(c_1, \ldots, c_n) \qquad \text{for } i = 1, 2, \ldots, n,$$

so the steady states are found by computing the zeros of F (in the case of a nonlinear system this may require numerical methods). Next, given the initial value of some solution, one wishes to analyze the behavior of this solution as $t \to \infty$. This is called the *asymptotic behavior* of the solution. Again, it may not be possible to find the solution explicitly, although by the general theory of differential equations a solution to the initial value problem is unique provided the functions F_i are differentiable. A steady state $y = c$ is called *globally asymptotically stable* if every solution tends to c as $t \to \infty$, i.e. for any solution $y(t)$ we have $\lim_{t \to \infty} y_i(t) = c_i$ for all $i = 1, 2, \ldots, n$.

In the case of the linear autonomous system $(*)$ the solutions form a vector space, so the only constant solution is the zero solution. The next exercise gives a *sufficient* condition for zero to be globally asymptotically stable and it gives one example of how the behavior of a linear system may be analyzed in terms of the eigenvalues of its coefficient matrix. Nonlinear systems can be approximated by linear systems in some neighborhood of a steady state by considering $y' = Ty$, where $T = \left(\dfrac{\partial F_i}{\partial y_j} \right)$ is the $n \times n$ Jacobian matrix of F evaluated at the steady state point. In this way the analysis of linear systems plays an important role in the local analysis of general autonomous systems.

59. Prove that the solution of $(*)$ given by $y_i(t) = 0$ for all $i \in \{1, \ldots, n\}$ (i.e. the zero solution) is globally asymptotically stable if all the eigenvalues of A have negative real parts. [For those unfamiliar with the behavior of the complex exponential function, assume all eigenvalues are real (hence are negative real numbers). Use the explicit nature of the solutions to show that they all tend to zero as $t \to \infty$.]

Part IV

FIELD THEORY AND GALOIS THEORY

The previous sections have developed the theory of some of the basic algebraic structures of groups, rings and fields. The next two chapters consider properties of fields, particularly fields which arise from trying to solve equations (such as the simple equation $x^2 + 1 = 0$), and fields which naturally arise in trying to perform "arithmetic" (adding, subtracting, multiplying and dividing). The elegant and beautiful Galois Theory relates the structure of *fields* to certain related *groups* and is one of the basic algebraic tools. Applications include solutions of classical compass and straightedge construction questions, finite fields and Abel's famous theorem on the insolvability (by radicals) of the general quintic polynomial.

13

Field Theory

13.1 BASIC THEORY OF FIELD EXTENSIONS

Recall that a field F is a commutative ring with identity in which every nonzero element has an inverse. Equivalently, the set $F^\times = F - \{0\}$ of nonzero elements of F is an abelian group under multiplication.

One of the first invariants associated with any field F is its *characteristic*, defined as follows: If 1_F denotes the identity of F, then F contains the elements 1_F, $1_F + 1_F$, $1_F + 1_F + 1_F$,... of the additive subgroup of F generated by 1_F, which may not all be distinct. For n a positive integer, let $n \cdot 1_F = 1_F + \ldots + 1_F$ (n times). Then two possibilities arise: either all the elements $n \cdot 1_F$ are distinct, or else $n \cdot 1_F = 0$ for some positive integer n.

Definition. The *characteristic* of a field F, denoted $\mathrm{ch}(F)$, is defined to be the smallest positive integer p such that $p \cdot 1_F = 0$ if such a p exists and is defined to be 0 otherwise.

It is easy to see that

$$n \cdot 1_F + m \cdot 1_F = (m + n) \cdot 1_F \qquad \text{and that}$$
$$(n \cdot 1_F)(m \cdot 1_F) = mn \cdot 1_F \tag{13.1}$$

for positive integers m and n. It follows that the characteristic of a field is either 0 or a *prime* p (hence the choice of p in the definition above), since if $n = ab$ is composite with $n \cdot 1_F = 0$, then $ab \cdot 1_F = (a \cdot 1_F)(b \cdot 1_F) = 0$ and since F is a field, one of $a \cdot 1_F$ or $b \cdot 1_F$ is 0, so the smallest such integer is necessarily a prime. It also follows that if $n \cdot 1_F = 0$, then n is divisible by p.

Proposition 1. The characteristic of a field F, $\mathrm{ch}(F)$, is either 0 or a prime p. For any $\alpha \in F$,

$$p \cdot \alpha = \underbrace{\alpha + \alpha + \ldots + \alpha}_{p \text{ times}} = 0.$$

Proof: Only the second statement has not been proved, and this follows immediately from the evident equality $p \cdot \alpha = p \cdot (1_F \alpha) = (p \cdot 1_F)(\alpha)$ in F.

Remark: This notion of a characteristic makes sense also for any integral domain and its characteristic will be the same as for its field of fractions.

Examples

(1) The fields $\mathbb{Q}$ and $\mathbb{R}$ both have characteristic 0: $\mathrm{ch}(\mathbb{Q}) = \mathrm{ch}(\mathbb{R}) = 0$. The integral domain $\mathbb{Z}$ also has characteristic 0.
(2) The (finite) field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ has characteristic p for any prime p.
(3) The integral domain $\mathbb{F}_p[x]$ of polynomials in the variable x with coefficients in the field $\mathbb{F}_p$ has characteristic p, as does its field of fractions $\mathbb{F}_p(x)$ (the field of rational functions in x with coefficients in $\mathbb{F}_p$).

If we define $(-n) \cdot 1_F = -(n \cdot 1_F)$ for positive n and $0 \cdot 1_F = 0$, then we have a natural ring homomorphism (by equation (1))

$$\begin{aligned} \varphi : \mathbb{Z} &\longrightarrow F \\ n &\longmapsto n \cdot 1_F \end{aligned}$$

and we can interpret the characteristic of F by noting that $\ker(\varphi) = \mathrm{ch}(F)\mathbb{Z}$. Taking the quotient by the kernel gives us an *injection* of either $\mathbb{Z}$ or $\mathbb{Z}/p\mathbb{Z}$ into F (depending on whether $\mathrm{ch}(F) = 0$ or $\mathrm{ch}(F) = p$). Since F is a field, we see that F contains a subfield isomorphic either to $\mathbb{Q}$ (the field of fractions of $\mathbb{Z}$) or to $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ (the field of fractions of $\mathbb{Z}/p\mathbb{Z}$) depending on the characteristic of F, and in either case is the smallest subfield of F containing 1_F (the field *generated* by 1_F in F).

Definition. The *prime subfield* of a field F is the subfield of F generated by the multiplicative identity 1_F of F. It is (isomorphic to) either $\mathbb{Q}$ (if $\mathrm{ch}(F) = 0$) or $\mathbb{F}_p$ (if $\mathrm{ch}(F) = p$).

Remark: We shall usually denote the identity 1_F of a field F simply by 1. Then in a field of characteristic p, one has $p \cdot 1 = 0$, frequently written simply $p = 0$ (for example, $2 = 0$ in a field of characterstic 2). It should be kept in mind, however, that this is a shorthand statement (the element "p" is really $p \cdot 1_F$ and is not a distinct element in F) to avoid confusion. This notation is computationally useful in light of the second statement in Proposition 1.

Examples

(1) The prime subfield of both $\mathbb{Q}$ and $\mathbb{R}$ is $\mathbb{Q}$.
(2) The prime subfield of the field $\mathbb{F}_p(x)$ is isomorphic to $\mathbb{F}_p$, given by the constant polynomials.

Definition. If K is a field containing the field F, then K is said to be an *extension field* (or simply an *extension*) of F, denoted K/F or by the diagram

$$\begin{array}{c} K \\ | \\ F \end{array}$$

In particular, every field F is an extension of its prime subfield. The field F is sometimes called the *base field* of the extension.

The notation K/F for a field extension is a shorthand for "K over F" and is not the quotient of K by F.

If K/F is any extension of fields, then the multiplication defined in K makes K into a *vector space* over F, in particular every field F can be considered as a vector space over its prime field.

Definition. The *degree* (or *relative degree* or *index*) of a field extension K/F, denoted $[K : F]$, is the dimension of K as a vector space over F (i.e. $[K : F] = \dim_F K$). The extension is said to be *finite* if $[K : F]$ is finite and is said to be *infinite* otherwise.

An important class of field extensions are those obtained by trying to solve equations over a given field F. For example, if $F = \mathbb{R}$ is the field of real numbers, then the simple equation $x^2 + 1 = 0$ does not have a solution in F. The question arises whether there is some larger field containing $\mathbb{R}$ in which this equation does have a solution, and it was this question that led Gauss to introduce the *complex numbers* $\mathbb{C} = \mathbb{R} + \mathbb{R}i$, where i is defined so that $i^2 + 1 = 0$. One then defines addition and multiplication in $\mathbb{C}$ by the usual rules familiar from elementary algebra and checks that in fact $\mathbb{C}$ so defined is a *field*, i.e. it is possible to find an inverse for every nonzero element of $\mathbb{C}$.

Given any field F and any polynomial $p(x) \in F[x]$ one can ask a similar question: does there exist an extension K of F containing a solution of the equation $p(x) = 0$ (i.e. containing a *root* of $p(x)$)? Note that we may assume here that the polynomial $p(x)$ is irreducible in $F[x]$ since a root of any factor of $p(x)$ is certainly a root of $p(x)$ itself. The answer is yes and follows almost immediately from our work on the polynomial ring $F[x]$. We first recall the following useful result on homomorphisms of fields (Corollary 10 of Chapter 7) which follows from the fact that the only ideals of a field F are 0 and F.

Proposition 2. Let $\varphi : F \to F'$ be a homomorphism of fields. Then φ is either identically 0 or is injective, so that the image of φ is either 0 or isomorphic to F.

Theorem 3. Let F be a field and let $p(x) \in F[x]$ be an irreducible polynomial. Then there exists a field K containing an isomorphic copy of F in which $p(x)$ has a root. Identifying F with this isomorphic copy shows that there exists an extension of F in which $p(x)$ has a root.

Proof: Consider the quotient

$$K = F[x]/(p(x))$$

of the polynomial ring $F[x]$ by the ideal generated by $p(x)$. Since by assumption $p(x)$ is an irreducible polynomial in the P.I.D. $F[x]$, the ideal $(p(x))$ is a *maximal* ideal. Hence K is actually a *field* (this is Proposition 12 of Chapter 7). The canonical projection π of $F[x]$ to the quotient $F[x]/(p(x))$ restricted to $F \subset F[x]$ gives a homomorphism $\varphi = \pi|_F : F \to K$ which is not identically 0 since it maps the identity 1 of F to the identity 1 of K. Hence by the proposition above, $\varphi(F) \cong F$ is an isomorphic copy of

F contained in K. If $\bar{x} = \pi(x)$ denotes the image of x in the quotient K, then

$$\begin{aligned} p(\bar{x}) &= \overline{p(x)} && \text{(since } \pi \text{ is a homomorphism)} \\ &= p(x) \pmod{p(x)} && \text{in } F[x]/(p(x)) \\ &= 0 && \text{in } F[x]/(p(x)) \end{aligned}$$

so that K does indeed contain a root of the polynomial $p(x)$. Since the field K naturally contains an isomorphic copy of F, we identify F with its isomorphic image in K and view F as a *subfield* of K. Then K is an extension of F in which the polynomial $p(x)$ has a root.

We shall use this result later to construct extensions of F containing *all* the roots of $p(x)$ (this is the notion of a *splitting field* and one of the central objects of interest in Galois theory).

To understand the field $K = F[x]/(p(x))$ constructed above more fully, it is useful to have a simple representation for the elements of this field. Since F is a subfield of K, we might in particular ask for a basis for K as a vector space over F.

Theorem 4. Let $p(x) \in F[x]$ be an irreducible polynomial of degree n over the field F and let K be the field $F[x]/(p(x))$. Let $\theta = x \bmod (p(x)) \in K$. Then the elements

$$1, \theta, \theta^2, \ldots, \theta^{n-1}$$

are a basis for K as a vector space over F, so the degree of the extension is n, i.e. $[K : F] = n$. Hence

$$K = \{a_0 + a_1\theta + a_2\theta^2 + \ldots + a_{n-1}\theta^{n-1} \mid a_0, a_1, \ldots, a_{n-1} \in F\}$$

consists of polynomials of degree $< n$ in θ.

Proof: Let $a(x) \in F[x]$ be any polynomial with coefficients in F. Since $F[x]$ is a Euclidean Domain (this is Theorem 3 of Chapter 9), we may divide $a(x)$ by $p(x)$:

$$a(x) = q(x)p(x) + r(x) \qquad q(x), r(x) \in F[x] \text{ with } \deg r(x) < n.$$

Since $q(x)p(x)$ lies in the ideal $(p(x))$, it follows that $a(x) \equiv r(x) \bmod (p(x))$, which shows that every residue class in $F[x]/(p(x))$ is represented by a polynomial of degree less than n. Hence the images $1, \theta, \theta^2, \ldots, \theta^{n-1}$ of $1, x, x^2, \ldots, x^{n-1}$ in the quotient *span* the quotient as a vector space over F. It remains to see that these elements are linearly independent, so form a *basis* for the quotient over F.

If the elements $1, \theta, \theta^2, \ldots, \theta^{n-1}$ were not linearly independent in K, then there would be a linear combination

$$b_0 + b_1\theta + b_2\theta^2 + \ldots + b_{n-1}\theta^{n-1} = 0$$

in K, with $b_0, b_1, \ldots, b_{n-1} \in F$, not all 0. This is equivalent to

$$b_0 + b_1x + b_2x^2 + \ldots + b_{n-1}x^{n-1} \equiv 0 \mod (p(x))$$

i.e.

$$p(x) \text{ divides } b_0 + b_1x + b_2x^2 + \ldots + b_{n-1}x^{n-1}$$

in $F[x]$. But this is impossible, since $p(x)$ is of degree n and the degree of the nonzero polynomial on the right is $< n$. This proves that $1, \theta, \theta^2, ..., \theta^{n-1}$ are a basis for K over F, so that $[K : F] = n$ by definition. The last statement of the theorem is clear.

This theorem provides an easy description of the elements of the field $F[x]/(p(x))$ as polynomials of degree $< n$ in θ where θ is an element (in K) with $p(\theta) = 0$. It remains only to see how to add and multiply elements written in this form. The addition in the quotient $F[x]/(p(x))$ is just usual addition of polynomials. The multiplication of polynomials $a(x)$ and $b(x)$ in the quotient $F[x]/(p(x))$ is performed by finding the product $a(x)b(x)$ in $F[x]$, then finding the representative of degree $< n$ for the coset $a(x)b(x) + (p(x))$ (as in the proof above) by dividing $a(x)b(x)$ by $p(x)$ and finding the remainder.

This can also be done easily in terms of θ as follows: We may suppose $p(x)$ is monic (since its roots and the ideal it generates do not change by multiplying by a constant), say $p(x) = x^n + p_{n-1}x^{n-1} + ... + p_1x + p_0$. Then in K, since $p(\theta) = 0$, we have

$$\theta^n = -(p_{n-1}\theta^{n-1} + ... + p_1\theta + p_0)$$

i.e. θ^n is a linear combination of lower powers of θ. Multiplying both sides by θ and replacing the θ^n on the right hand side by these lower powers again, we see that also θ^{n+1} is a polynomial of degree $< n$ in θ. Similarly, any positive power of θ can be written as a polynomial of degree $< n$ in θ, hence *any* polynomial in θ can be written as a polynomial of degree $< n$ in θ. Multiplication in K is now easily performed: one simply writes the product of two polynomials of degree $< n$ in θ as another polynomial of degree $< n$ in θ.

We summarize this as:

Corollary 5. Let K be as in Theorem 4, and let $a(\theta), b(\theta) \in K$ be two polynomials of degree $< n$ in θ. Then addition in K is defined simply by usual polynomial addition and multiplication in K is defined by

$$a(\theta)b(\theta) = r(\theta)$$

where $r(x)$ is the remainder (of degree $< n$) obtained after dividing the polynomial $a(x)b(x)$ by $p(x)$ in $F[x]$.

By the results proved above, this definition of addition and multiplication on the polynomials of degree $< n$ in θ make K into a *field*, so that one can also *divide* by nonzero elements as well, which is not so immediately obvious from the definitions of the operations.

It is also important in Theorem 4 that the polynomial $p(x)$ be *irreducible* over F. In general the addition and multiplication in Corollary 5 (which can be defined in the same way for any polynomial $p(x)$) do *not* make the polynomials of degree $< n$ in θ into a field if $p(x)$ is not irreducible. In fact, this set is not even an integral domain in general (its structure is given by Proposition 16 of Chapter 9). To describe the *field* containing a root θ of a general polynomial $f(x)$ over F, $f(x)$ is factored into irreducibles in $F[x]$ and the results above are applied to an irreducible factor $p(x)$ of $f(x)$ having θ as a root. We shall consider this more in the following sections.

Examples

(1) If we apply this construction to the special case $F = \mathbb{R}$ and $p(x) = x^2 + 1$ then we obtain the field

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

which is an extension of degree 2 of $\mathbb{R}$ in which $x^2 + 1$ has a root. The elements of this field are of the form $a + b\theta$ for $a, b \in \mathbb{R}$. Addition is defined by

$$(a + b\theta) + (c + d\theta) = (a + c) + (b + d)\theta. \tag{13.2a}$$

To multiply we use the fact that $\theta^2 + 1 = 0$, i.e. $\theta^2 = -1$ in K. (Alternatively, note that -1 is also the remainder when x^2 is divided by $x^2 + 1$ in $\mathbb{R}[x]$.) Then

$$\begin{aligned}(a + b\theta)(c + d\theta) &= ac + (ad + bc)\theta + bd\theta^2 \\ &= ac + (ad + bc)\theta + bd(-1) \\ &= (ac - bd) + (ad + bc)\theta \quad .\end{aligned} \tag{13.2b}$$

These are, up to changing θ to i, the formulas for adding and multiplying in $\mathbb{C}$. Put another way, the map

$$\begin{aligned}\varphi : \mathbb{R}[x]/(x^2 + 1) &\longrightarrow \mathbb{C} \\ a + bx &\mapsto a + bi\end{aligned}$$

is a homomorphism. Since it is bijective (as a map of vector spaces over the reals, for example), it is an isomorphism. Notice that instead of taking the existence of $\mathbb{C}$ for granted (along with the fairly tedious verification that it is in fact a field), we could have *defined* $\mathbb{C}$ by this isomorphism. Then the fact that it is a field is a consequence of Theorem 4.

(2) Take now $F = \mathbb{Q}$ to be the field of rational numbers and again take $p(x) = x^2 + 1$ (still irreducible over $\mathbb{Q}$, of course). Then the same construction, with the same addition and multiplication formulas as (2a) and (2b) above, except that now a and b are elements of $\mathbb{Q}$, defines a field extension $\mathbb{Q}(i)$ of $\mathbb{Q}$ of degree 2 containing a root i of $x^2 + 1$.

(3) Take $F = \mathbb{Q}$ and $p(x) = x^2 - 2$, irreducible over $\mathbb{Q}$ by Eisenstein's Criterion, for example. Then we obtain a field extension of $\mathbb{Q}$ of degree 2 containing a square root θ of 2, denoted $\mathbb{Q}(\theta)$. If we denote θ by $\sqrt{2}$, the elements of this field are of the form

$$a + b\sqrt{2}, \qquad a, b \in \mathbb{Q}$$

with addition defined by

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

and multiplication defined by

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

(4) Let $F = \mathbb{Q}$ and $p(x) = x^3 - 2$, irreducible again by Eisenstein. Denoting a root of $p(x)$ by θ, we obtain the field

$$\mathbb{Q}[x]/(x^3 - 2) \cong \{a + b\theta + c\theta^2 \mid a, b, c \in \mathbb{Q}\}$$

with $\theta^3 = 2$, an extension of degree 3. To find the inverse of, say, $1 + \theta$ in this field, we can proceed as follows: By the Euclidean Algorithm in $\mathbb{Q}[x]$ there are polynomials

$a(x)$ and $b(x)$ with

$$a(x)(1+x)+b(x)(x^3-2)=1$$

(since $p(x)=x^3-2$ is irreducible, it is relatively prime to every polynomial of smaller degree). In the quotient field this equation implies that $a(\theta)$ is the inverse of $1+\theta$. In this case, a simple computation shows that we can take $a(x)=\frac{1}{3}(x^2-x+1)$ (and $b(x)=-\frac{1}{3}$), so that

$$(1+\theta)^{-1}=\frac{\theta^2-\theta+1}{3}.$$

(5) In general, if $\theta \in K$ is a root of the irreducible polynomial

$$p(x)=p_n x^n+p_{n-1}x^{n-1}+\ldots+p_1 x+p_0$$

we can compute $\theta^{-1} \in K$ from

$$\theta(p_n\theta^{n-1}+p_{n-1}\theta^{n-2}+\ldots+p_1)=-p_0$$

namely

$$\theta^{-1}=\frac{-1}{p_0}(p_n\theta^{n-1}+p_{n-1}\theta^{n-2}+\ldots+p_1)\in K$$

(note that $p_0 \neq 0$ since $p(x)$ is irreducible).

Remark: Determining inverses in extensions of this type may be familiar from elementary algebra in the case of $\mathbb{C}$ or Example 3 under the name "rationalizing denominators". The last two examples indicate a procedure which is much more general than the *ad hoc* procedures of elementary algebra.

(6) Take $F=\mathbb{F}_2$, the finite field with two elements, and $p(x)=x^2+x+1$, which we have previously checked is irreducible over $\mathbb{F}_2$. Here we obtain a degree 2 extension of $\mathbb{F}_2$

$$\mathbb{F}_2[x]/(x^2+x+1)\cong\{a+b\theta \mid a,b\in\mathbb{F}_2\}$$

where $\theta^2=-\theta-1=\theta+1$. Multiplication in this field $\mathbb{F}_2(\theta)$ (which contains four elements) is defined by

$$\begin{aligned}(a+b\theta)(c+d\theta)&=ac+(ad+bc)\theta+bd\theta^2\\&=ac+(ad+bc)\theta+bd(\theta+1)\\&=(ac+bd)+(ad+bc+bd)\theta.\end{aligned}$$

(7) Let $F=k(t)$ be the field of rational functions in the variable t over a field k (for example, $k=\mathbb{Q}$ or $k=\mathbb{F}_p$). Let $p(x)=x^2-t\in F[x]$. Then $p(x)$ is irreducible (it is Eisenstein at the prime (t) in $k[t]$). If we denote a root by θ, the corresponding degree 2 field extension $F(\theta)$ consists of the elements

$$\{a(t)+b(t)\theta \mid a(t),b(t)\in F\}$$

where the coefficients $a(t)$ and $b(t)$ are rational functions in t with coefficients in k and where $\theta^2=t$.

Suppose F is a subfield of a field K and $\alpha\in K$ is an element of K. Then the collection of subfields of K containing both F and α is nonempty (K is such a field,

for example). Since the intersection of subfields is again a subfield, it follows that there is a unique minimal subfield of K containing both F and α (the intersection of all subfields with this property). Similar remarks apply if α is replaced by a collection $\alpha, \beta, \ldots$ of elements of K.

Definition. Let K be an extension of the field F and let $\alpha, \beta, \ldots \in K$ be a collection of elements of K. Then the smallest subfield of K containing both F and the elements $\alpha, \beta, \ldots$, denoted $F(\alpha, \beta, \ldots)$ is called the field *generated by* $\alpha, \beta, \ldots$ *over* F.

Definition. If the field K is generated by a single element α over F, $K = F(\alpha)$, then K is said to be a *simple* extension of F and the element α is called a *primitive element* for the extension.

We shall later characterize which extensions of a field F are simple. In particular we shall prove that every finite extension of a field of characteristic 0 is a simple extension.

The connection between the simple extension $F(\alpha)$ generated by α over F where α is a root of some irreducible polynomial $p(x)$ and the field constructed in Theorem 4 is provided by the following:

Theorem 6. Let F be a field and let $p(x) \in F[x]$ be an irreducible polynomial. Suppose K is an extension field of F containing a root α of $p(x)$: $p(\alpha) = 0$. Let $F(\alpha)$ denote the subfield of K generated over F by α. Then

$$F(\alpha) \cong F[x]/(p(x)).$$

Remark: This theorem says that *any* field over F in which $p(x)$ contains a root contains a subfield isomorphic to the extension of F constructed in Theorem 3 and that this field is (up to isomorphism) the smallest extension of F containing such a root. The difference between this result and Theorem 3 is that Theorem 6 *assumes* the existence of a root α of $p(x)$ in some field K and the major point of Theorem 3 is *proving* that there exists such an extension field K.

Proof: There is a natural homomorphism

$$\begin{aligned} \varphi : F[x] &\longrightarrow F(\alpha) \subseteq K \\ a(x) &\longmapsto a(\alpha) \end{aligned}$$

obtained by mapping F to F by the identity map and sending x to α and then extending so that the map is a ring homomorphism (i.e. the polynomial $a(x)$ in x maps to the polynomial $a(\alpha)$ in α). Since $p(\alpha) = 0$ by assumption, the element $p(x)$ is in the kernel of φ, so we obtain an induced homomorphism (also denoted φ):

$$\varphi : F[x]/(p(x)) \longrightarrow F(\alpha).$$

But since $p(x)$ is irreducible, the quotient on the left is a *field,* and φ is not the 0 map (it is the identity on F, for example), hence φ is an isomorphism of the field on the left

with its image. Since this image is then a subfield of $F(\alpha)$ containing F and containing α, by the definition of $F(\alpha)$ the map must be surjective, proving the theorem.

Combined with Corollary 5, this determines the field $F(\alpha)$ when α is a root of an irreducible polynomial $p(x)$:

Corollary 7. Suppose in Theorem 6 that $p(x)$ is of degree n. Then

$$F(\alpha) = \{a_0 + a_1\alpha + a_2\alpha^2 + \ldots + a_{n-1}\alpha^{n-1} \mid a_0, a_1, \ldots, a_{n-1} \in F\} \subseteq K.$$

Describing fields generated by more than one element is more complicated and we shall return to this question in the following section.

Examples

(1) In Example 3 above, we have determined the field $\mathbb{Q}(\sqrt{2})$ generated over $\mathbb{Q}$ by the element $\sqrt{2} \in \mathbb{R}$, having suggestively denoted the abstract solution θ of the equation $x^2 - 2 = 0$ by the symbol $\sqrt{2}$, which has an independent meaning in the field $\mathbb{R}$ (namely the *positive* square root of 2 in $\mathbb{R}$).

(2) The equation $x^2 - 2 = 0$ has another solution in $\mathbb{R}$, namely $-\sqrt{2}$, the *negative* square root of 2 in $\mathbb{R}$. The field generated over $\mathbb{Q}$ by this solution consists of the elements $\{a + b(-\sqrt{2}) \mid a, b \in \mathbb{Q}\}$, and is again isomorphic to the field in Example 3 above (hence also isomorphic to the field just considered, the isomorphism given explicitly by $a + b\sqrt{2} \mapsto a - b\sqrt{2}$). As a subset of $\mathbb{R}$ this is the same set of elements as in Example 1.

(3) Similarly, if we use the symbol $\sqrt[3]{2}$ to denote the (positive) cube root of 2 in $\mathbb{R}$, then the field generated by $\sqrt[3]{2}$ over $\mathbb{Q}$ in $\mathbb{R}$ consists of the elements

$$\{a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \mid a, b, c \in \mathbb{Q}\}$$

and is isomorphic to the field constructed in Example 4 above.

(4) The equation $x^3 - 2 = 0$ has no further solutions in $\mathbb{R}$, but there are two additional solutions in $\mathbb{C}$ given by $\sqrt[3]{2}(\frac{-1 + i\sqrt{3}}{2})$ and $\sqrt[3]{2}(\frac{-1 - i\sqrt{3}}{2})$ ($\sqrt{3}$ denoting the positive real square root of 3) as can easily be checked. The fields generated by either of these two elements over $\mathbb{Q}$ are subfields of $\mathbb{C}$ (but not of $\mathbb{R}$) and are both isomorphic to the field constructed in the previous example (and to Example 4 earlier).

As Theorem 6 indicates, the roots of an irreducible polynomial $p(x)$ are *algebraically indistinguishable* in the sense that the fields obtained by adjoining any root of an irreducible polynomial are isomorphic. In the last two examples above, the fields obtained by adjoining one of the three possible (complex) roots of $x^3 - 2 = 0$ to $\mathbb{Q}$ were all algebraically isomorphic. The fields were distinguished not by their algebraic properties, but by whether their elements were *real*, which involves *continuous* operations.

The fact that different roots of the same irreducible polynomial have the same algebraic properties can be extended slightly, as follows:

Let $\varphi : F \xrightarrow{\sim} F'$ be an isomorphism of fields. The map φ induces a ring isomorphism (also denoted φ)

$$\varphi : F[x] \xrightarrow{\sim} F'[x]$$

defined by applying φ to the coefficients of a polynomial in $F[x]$. Let $p(x) \in F[x]$ be an irreducible polynomial and let $p'(x) \in F'[x]$ be the polynomial obtained by applying the map φ to the coefficients of $p(x)$, i.e. the image of $p(x)$ under φ. The isomorphism φ maps the maximal ideal $(p(x))$ to the ideal $(p'(x))$, so this ideal is also maximal, which shows that $p'(x)$ is also irreducible in $F'[x]$. The following theorem shows that the fields obtained by adjoining a root of $p(x)$ to F and a root of $p'(x)$ to F' have the same algebraic structure (i.e. are isomorphic):

Theorem 8. Let $\varphi : F \xrightarrow{\sim} F'$ be an isomorphism of fields. Let $p(x) \in F[x]$ be an irreducible polynomial and let $p'(x) \in F'[x]$ be the irreducible polynomial obtained by applying the map φ to the coefficients of $p(x)$. Let α be a root of $p(x)$ (in some extension of F) and let β be a root of $p'(x)$ (in some extension of F'). Then there is an isomorphism

$$\begin{aligned} \sigma : F(\alpha) &\xrightarrow{\sim} F'(\beta) \\ \alpha &\mapsto \beta \end{aligned}$$

mapping α to β and extending φ, i.e. such that σ restricted to F is the isomorphism φ.

Proof: As noted above, the isomorphism φ induces a natural isomorphism from $F[x]$ to $F'[x]$ which maps the maximal ideal $(p(x))$ to the maximal ideal $(p'(x))$. Taking the quotients by these ideals, we obtain an isomorphism of fields

$$F[x]/(p(x)) \xrightarrow{\sim} F'[x]/(p'(x)).$$

By Theorem 6 the field on the left is isomorphic to $F(\alpha)$ and by the same theorem the field on the right is isomorphic to $F'(\beta)$. Composing these isomorphisms, we obtain the isomorphism σ. It is clear that the restriction of this isomorphism to F is φ, completing the proof.

This extension theorem will be of considerable use when we consider Galois Theory later. It can be represented pictorially by the diagram:

$$\begin{array}{cccc} \sigma : & F(\alpha) & \xrightarrow{\sim} & F'(\beta) \\ & | & & | \\ \varphi : & F & \xrightarrow{\sim} & F' \end{array}$$

EXERCISES

1. Show that $p(x) = x^3 + 9x + 6$ is irreducible in $\mathbb{Q}[x]$. Let θ be a root of $p(x)$. Find the inverse of $1 + \theta$ in $\mathbb{Q}(\theta)$.
2. Show that $x^3 - 2x - 2$ is irreducible over $\mathbb{Q}$ and let θ be a root. Compute $(1+\theta)(1+\theta+\theta^2)$ and $\dfrac{1+\theta}{1+\theta+\theta^2}$ in $\mathbb{Q}(\theta)$.
3. Show that $x^3 + x + 1$ is irreducible over $\mathbb{F}_2$ and let θ be a root. Compute the powers of θ in $\mathbb{F}_2(\theta)$.
4. Prove directly that the map $a + b\sqrt{2} \mapsto a - b\sqrt{2}$ is an isomorphism of $\mathbb{Q}(\sqrt{2})$ with itself.
5. Suppose α is a rational root of a monic polynomial in $\mathbb{Z}[x]$. Prove that α is an integer.

6. Show that if α is a root of $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ then $a_n\alpha$ is a root of the monic polynomial $x^n + a_n a_{n-1}x^{n-1} + \cdots + a_n^{n-1}a_1 x + a_n^n a_0$.
7. Prove that $x^3 - nx + 2$ is irreducible for $n \neq -1, 3, 5$.
8. Prove that $x^5 - ax - 1 \in \mathbb{Z}[x]$ is irreducible unless $a = 0, 2$ or -1. The first two correspond to linear factors, the third corresponds to the factorization $(x^2 - x + 1)(x^3 + x^2 - 1)$.

13.2 ALGEBRAIC EXTENSIONS

Let F be a field and let K be an extension of F.

Definition. The element $\alpha \in K$ is said to be *algebraic* over F if α is a root of some nonzero polynomial $f(x) \in F[x]$. If α is not algebraic over F (i.e. is not the root of any nonzero polynomial with coefficients in F) then α is said to be *transcendental* over F. The extension K/F is said to be *algebraic* if every element of K is algebraic over F.

Note that if α is algebraic over a field F then it is algebraic over any extension field L of F (if $f(x)$ having α as a root has coefficients in F then it also has coefficients in L).

Proposition 9. Let α be algebraic over F. Then there is a unique monic irreducible polynomial $m_{\alpha,F}(x) \in F[x]$ which has α as a root. A polynomial $f(x) \in F[x]$ has α as a root if and only if $m_{\alpha,F}(x)$ divides $f(x)$ in $F[x]$.

Proof: Let $g(x) \in F[x]$ be a polynomial of minimal degree having α as a root. Multiplying $g(x)$ by a constant, we may assume $g(x)$ is monic. Suppose $g(x)$ were reducible in $F[x]$, say $g(x) = a(x)b(x)$ with $a(x), b(x) \in F[x]$ both of degree smaller than the degree of $g(x)$. Then $g(\alpha) = a(\alpha)b(\alpha)$ in K, and since K is a field, either $a(\alpha) = 0$ or $b(\alpha) = 0$, contradicting the minimality of the degree of $g(x)$. It follows that $g(x)$ is a monic irreducible polynomial having α as a root. Suppose now that $f(x) \in F[x]$ is any polynomial having α as a root. By the Euclidean Algorithm in $F[x]$ there are polynomials $q(x), r(x) \in F[x]$ such that

$$f(x) = q(x)g(x) + r(x) \qquad \text{with} \quad \deg r(x) < \deg g(x).$$

Then $f(\alpha) = q(\alpha)g(\alpha) + r(\alpha)$ in K and since α is a root of both $f(x)$ and $g(x)$, we obtain $r(\alpha) = 0$, which contradicts the minimality of $g(x)$ unless $r(x) = 0$. Hence $g(x)$ divides any polynomial $f(x)$ in $F[x]$ having α as a root and in particular would divide any other monic irreducible polynomial in $F[x]$ having α as a root. This proves that $m_{\alpha,F}(x) = g(x)$ is unique and completes the proof of the proposition.

Corollary 10. If L/F is an extension of fields and α is algebraic over both F and L, then $m_{\alpha,L}(x)$ divides $m_{\alpha,F}(x)$ in $L[x]$.

Proof: This is immediate from the second statement in Proposition 9 applied to L, since $m_{\alpha,F}(x)$ is a polynomial in $L[x]$ having α as a root.

Definition. The polynomial $m_{\alpha,F}(x)$ (or just $m_\alpha(x)$ if the field F is understood) in Proposition 9 is called the *minimal polynomial* for α over F. The *degree* of $m_\alpha(x)$ is called the *degree* of α.

Note that by the proposition, a monic polynomial over F with α as a root is the minimal polynomial for α over F if and only if it is irreducible over F.

Proposition 11. Let α be algebraic over the field F and let $F(\alpha)$ be the field generated by α over F. Then

$$F(\alpha) \cong F[x]/(m_\alpha(x))$$

so that in particular

$$[F(\alpha) : F] = \deg\ m_\alpha(x) = \deg \alpha,$$

i.e. the degree of α over F is the degree of the extension it generates over F.

Proof: This follows immediately from Theorem 6.

Examples

(1) The minimal polynomial for $\sqrt{2}$ over $\mathbb{Q}$ is $x^2 - 2$ and $\sqrt{2}$ is of degree 2 over $\mathbb{Q}$: $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$.

(2) The minimal polynomial for $\sqrt[3]{2}$ over $\mathbb{Q}$ is $x^3 - 2$ and $\sqrt[3]{2}$ is of degree 3 over $\mathbb{Q}$: $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$.

(3) Similarly, for any $n > 1$, the polynomial $x^n - 2$ is irreducible over $\mathbb{Q}$ since it is Eisenstein. Denoting a root of this polynomial by $\sqrt[n]{2}$ (where as usual we reserve this symbol to denote the *positive* n^{th} root of 2 if we want to view this root as an element of $\mathbb{R}$, and where the symbol denotes any one of the algebraically indistinguishable abstract solutions in general), we have $[\mathbb{Q}(\sqrt[n]{2}) : \mathbb{Q}] = n$.

(4) The minimal polynomial and the degree of an element α depend on the base field. For example, over $\mathbb{R}$, the element $\sqrt[n]{2}$ is of degree *one*, with minimal polynomial $m_{\sqrt[n]{2},\mathbb{R}}(x) = x - \sqrt[n]{2}$.

(5) Consider the polynomial $p(x) = x^3 - 3x - 1$ over $\mathbb{Q}$, which is irreducible over $\mathbb{Q}$ since it is a cubic which has no rational root (cf. Proposition 11 of Chapter 9), hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$ for any root α of $p(x)$. For future reference we note that a quick sketch of the graph of this function over the real numbers shows that the graph crosses the x-axis precisely once in the interval [0,2], i.e. there is precisely one real number α, $0 < \alpha < 2$ satisfying $\alpha^3 - 3\alpha - 1 = 0$.

Proposition 12. The element α is algebraic over F if and only if the simple extension $F(\alpha)/F$ is finite. More precisely, if α is an element of an extension of degree n over F then α satisfies a polynomial of degree at most n over F and if α satisfies a polynomial of degree n over F then the degree of $F(\alpha)$ over F is at most n.

Proof: If α is algebraic over F, then the degree of the extension $F(\alpha)/F$ is the degree of the minimal polynomial for α over F, hence the extension is finite, of degree $\leq n$ if α satisfies a polynomial of degree n. Conversely, suppose α is an element of an extension of degree n over F (for example, if $[F(\alpha) : F] = n$). Then the $n + 1$ elements

$$1, \alpha, \alpha^2, \ldots, \alpha^n$$

of $F(\alpha)$ are linearly dependent over F, say

$$b_0 + b_1\alpha + b_2\alpha^2 + \ldots + b_n\alpha^n = 0$$

with $b_0, b_1, b_2, \ldots, b_n \in F$ not all 0. Hence α is the root of a nonzero polynomial with coefficients in F (of degree $\leq n$), which proves α is algebraic over F and also proves the second statement of the proposition.

Corollary 13. If the extension K/F is finite, then it is algebraic.

Proof: If $\alpha \in K$, then the subfield $F(\alpha)$ is in particular a subspace of the vector space K over F. Hence $[F(\alpha) : F] \leq [K : F]$ and so α is algebraic over F by the proposition.

Remark: We shall prove below a sort of converse to this result (Theorem 17), but note that there are infinite algebraic extensions (we shall have an example later), so the literal converse of this corollary is not true.

Example: (Quadratic Extensions over fields of characteristic $\neq 2$)

Let F be a field of characteristic $\neq 2$ (for example, any field of characteristic 0, such as $\mathbb{Q}$) and let K be an extension of F of degree 2, $[K : F] = 2$. Let α be any element of K not contained in F. By the proposition above, α satisfies an equation of degree at most 2 over F. This equation cannot be of degree 1, since α is not an element of F by assumption. It follows that the minimal polynomial of α is a monic quadratic

$$m_\alpha(x) = x^2 + bx + c \qquad b, c \in F.$$

Since $F \subset F(\alpha) \subseteq K$ and $F(\alpha)$ is already a vector space over F of dimension 2, we have $K = F(\alpha)$.

The roots of this quadratic equation can be determined by the quadratic formula, which is valid over any field of characteristic $\neq 2$ (the formula is obtained as in elementary algebra by completing the square):

$$\alpha = \frac{-b \pm \sqrt{b^2 - 4c}}{2}$$

(the reason for requiring the characteristic of F not be 2 is that we must divide by 2). Here $b^2 - 4c$ is not a square in F since α is not an element of F and the symbol $\sqrt{b^2 - 4c}$ denotes a root of the equation $x^2 - (b^2 - 4c) = 0$ in K (see the end of the next paragraph). Note that here there is no natural choice of one of the roots analogous to choosing the *positive* square root of 2 in $\mathbb{R}$ — the roots are algebraically indistinguishable.

Now $F(\alpha) = F(\sqrt{b^2 - 4c})$ as follows: by the formula above, α is an element of the field on the right, hence $F(\alpha) \subseteq F(\sqrt{b^2 - 4c})$. Conversely, $\sqrt{b^2 - 4c} = \mp(b + 2\alpha)$ shows that $\sqrt{b^2 - 4c}$ is an element of $F(\alpha)$, which gives the reverse inclusion $F(\sqrt{b^2 - 4c}) \subseteq F(\alpha)$ (and incidentally shows that the equation $x^2 - (b^2 - 4c) = 0$ does have a solution in K).

It follows that any extension K of F of degree 2 is of the form $F(\sqrt{D})$ where D is an element of F which is not a square in F, and conversely, every such extension is an extension of degree 2 of F. For this reason, extensions of degree 2 of a field F are called *quadratic* extensions of F.

Suppose that F is a subfield of a field K which in turn is a subfield of a field L. Then there are three associated extension degrees – the dimension of K and L as vector spaces over F, and the dimension of L as a vector space over K.

Theorem 14. Let $F \subseteq K \subseteq L$ be fields. Then

$$[L : F] = [L : K][K : F]$$

i.e. extension degrees are multiplicative, where if one side of the equation is infinite, the other side is also infinite. Pictorially:

$$\overbrace{\underbrace{F \quad \subseteq \quad K}_{[K:F]} \quad \underbrace{\subseteq \quad L}_{[L:K]}}^{[L:F]}$$

Proof: Suppose first that $[L : K] = m$ and $[K : F] = n$ are finite. Let $\alpha_1, \alpha_2, ..., \alpha_m$ be a basis for L over K and let $\beta_1, \beta_2, ..., \beta_n$ be a basis for K over F. Then every element of L can be written as a linear combination

$$a_1\alpha_1 + a_2\alpha_2 + ... + a_m\alpha_m$$

where $a_1, a_2, ..., a_m$ are elements of K, hence are linear combinations of $\beta_1, \beta_2, ..., \beta_n$:

$$a_i = b_{i1}\beta_1 + b_{i2}\beta_2 + ... + b_{in}\beta_n \qquad i = 1, 2, ..., m \tag{13.3}$$

where the b_{ij} are elements of F. Substituting these expressions in for the coefficients a_i above, we see that every element of L can be written as a linear combination

$$\sum_{\substack{i=1,2,...,m \\ j=1,2,...,n}} b_{ij}\alpha_i\beta_j$$

of the mn elements $\alpha_i\beta_j$ with coefficients in F. Hence these elements *span* L as a vector space over F.

Suppose now that we had a linear relation in L

$$\sum_{\substack{i=1,2,...,m \\ j=1,2,...,n}} b_{ij}\alpha_i\beta_j = 0$$

with coefficients b_{ij} in F. Then defining the elements $a_i \in K$ by equation (3) above, this linear relation could be written

$$a_1\alpha_1 + a_2\alpha_2 + ... + a_m\alpha_m = 0.$$

Since the α_i are a basis for L over K, it follows that all the coefficients a_i, $i = 1, 2, ..., m$ must be 0, i.e. that

$$b_{i1}\beta_1 + b_{i2}\beta_2 + ... + b_{in}\beta_n = 0 \qquad i = 1, 2, ..., m$$

in K. Since now the $\beta_j, j = 1, 2, ..., n$ are a basis for K over F, this implies $b_{ij} = 0$ for all i and j. Hence the elements $\alpha_i\beta_j$ are linearly independent over F, so are a basis for L over F and $[L : F] = mn = [L : K][K : F]$, as claimed.

If $[K : F]$ is infinite, then there are infinitely many elements of K, hence of L, which are linearly independent over F, so that $[L : F]$ is also infinite. Similarly, if $[L : K]$ is infinite, there are infinitely many elements of L linearly independent over K, so certainly linearly independent over F, so again $[L : F]$ is infinite. Finally, if $[L : K]$ and $[K : F]$ are both finite, then the proof above shows $[L : F]$ is finite, so that $[L : F]$ infinite implies at least one of $[L : K]$ and $[K : F]$ is infinite, completing the proof.

Remark: Note the similarity of this result with the result on group orders proved in Part I. As with diagrams involving groups we shall frequently indicate the relative degrees of extensions in field diagrams.

The multiplicativity of extension degrees is extremely useful in computations. A particular application is:

Corollary 15. Suppose L/F is a finite extension and let K be any subfield of L containing F, $F \subseteq K \subseteq L$. Then $[K : F]$ divides $[L : F]$.

Proof: This is immediate.

Examples

(1) The element $\sqrt{2}$ is not contained in the field $\mathbb{Q}(\alpha)$ where α is the real root of $x^3 - 3x - 1$ between 0 and 2, since we have already determined $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$ and $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$ and 2 does not divide 3. Note that it is not so easy to prove directly that $\sqrt{2}$ cannot be written as a rational linear combination of $1, \alpha, \alpha^2$.

(2) Let as usual $\sqrt[6]{2}$ denote the positive real 6th root of 2. Then $[\mathbb{Q}(\sqrt[6]{2}) : \mathbb{Q}] = 6$. Since $(\sqrt[6]{2})^3 = \sqrt{2}$ we have $\mathbb{Q}(\sqrt{2}) \subset \mathbb{Q}(\sqrt[6]{2})$ and by the multiplicativity of extension degrees, $[\mathbb{Q}(\sqrt[6]{2}) : \mathbb{Q}(\sqrt{2})] = 3$. This gives us the field diagram

$$\overbrace{\underbrace{\mathbb{Q} \quad \subset \quad \mathbb{Q}(\sqrt{2})}_{2} \quad \subset \quad \mathbb{Q}(\sqrt[6]{2})}^{6}$$

$$\underbrace{\qquad\qquad}_{3}$$

In particular, this shows that the minimal polynomial for $\sqrt[6]{2}$ over $\mathbb{Q}(\sqrt{2})$ is of degree 3. It is therefore the polynomial $x^3 - \sqrt{2}$. Note that showing directly that this polynomial is irreducible over $\mathbb{Q}(\sqrt{2})$ is not completely trivial.

By Theorem 14 a finite extension of a finite extension is finite. The next results use this to show that an extension generated by a finite number of algebraic elements is finite (extending Proposition 12).

Definition. An extension K/F is *finitely generated* if there are elements $\alpha_1, \alpha_2, \ldots, \alpha_k \in K$ such that $K = F(\alpha_1, \alpha_2, \ldots, \alpha_k)$.

Recall that the field generated over F by a collection of elements in a field K is the smallest subfield of K containing these elements and F. The next lemma will

show that for finitely generated extensions this field can be obtained recursively by a series of simple extensions.

Lemma 16. $F(\alpha, \beta) = (F(\alpha))(\beta)$, i.e. the field generated over F by α and β is the field generated by β over the field $F(\alpha)$ generated by α.

Proof: This follows by the minimality of the fields in question. The field $F(\alpha, \beta)$ contains F and α, hence contains the field $F(\alpha)$, and since it also contains β, we have the inclusion $(F(\alpha))(\beta) \subseteq F(\alpha, \beta)$ by the minimality of the field $(F(\alpha))(\beta)$. Since the field $(F(\alpha))(\beta)$ contains F, α and β, by the minimality of $F(\alpha, \beta)$ we have the reverse inclusion $F(\alpha, \beta) \subseteq (F(\alpha))(\beta)$, which proves the lemma.

By the lemma we have

$$K = F(\alpha_1, \alpha_2, ..., \alpha_k) = (F(\alpha_1, \alpha_2, ..., \alpha_{k-1}))(\alpha_k)$$

and so by iterating, we see that K is obtained by taking the field F_1 generated over F by α_1, then the field F_2 generated *over* F_1 (this is important) by α_2, and so on, with $F_k = K$. This gives a sequence of fields:

$$F = F_0 \subseteq F_1 \subseteq F_2 \subseteq ... \subseteq F_k = K$$

where

$$F_{i+1} = F_i(\alpha_{i+1}) \qquad i = 0, 1, ..., k-1.$$

Suppose now that the elements $\alpha_1, \alpha_2, ..., \alpha_k$ are algebraic over F of degrees $n_1, n_2, ..., n_k$ (so a priori are algebraic over any extension of F). Then the extensions in this sequence are simple extensions of the type considered in Proposition 11. The relative extension degree $[F_{i+1} : F_i]$ is equal to the degree of the minimal polynomial of α_{i+1} over F_i, which is at most n_{i+1} (and equals n_{i+1} if and only if the minimal polynomial of α_{i+1} over F remains irreducible over F_i). By the multiplicativity of extension degrees, we see that

$$[K : F] = [F_k : F_{k-1}][F_{k-1} : F_{k-2}]...[F_1 : F_0]$$

is also finite, and $\leq n_1 n_2 ... n_k$.

This also gives a description of the elements of $F(\alpha_1, \alpha_2, ..., \alpha_k)$. For simplicity, consider the case of the field $F(\alpha, \beta)$ where α and β are algebraic over F. Then the elements of this field are of the form

$$b_0 + b_1\beta + b_2\beta^2 + ... + b_{d-1}\beta^{d-1}$$

where $d = [F(\alpha)(\beta) : F(\alpha)]$ is the degree of β over $F(\alpha)$ (which may be strictly smaller than the degree of β over F, recall), and where the coefficients $b_0, b_1, ..., b_d$ are elements of $F(\alpha)$. The coefficients $b_i \in F(\alpha)$, $i = 0, ..., d-1$, are of the form

$$a_{0i} + a_{1i}\alpha + a_{2i}\alpha^2 + ... + a_{n\text{-}1i}\alpha^{n-1}$$

where $n = [F(\alpha) : F]$ is the degree of α over F and the a_{ij} are elements of F. Hence the elements of $F(\alpha, \beta)$ are of the form

$$\sum_{\substack{i=0,1,\ldots,n-1 \\ j=0,1,\ldots,d-1}} a_{ij}\alpha^i\beta^j \qquad a_{ij} \in F.$$

Since $[F(\alpha, \beta) : F] = [F(\alpha, \beta) : F(\alpha)][F(\alpha) : F] = dn$, the elements $\alpha^i\beta^j$ are in fact an F basis for $F(\alpha, \beta)$.

In practice the field $F(\alpha)$ generated by the algebraic α is obtained by adjoining the element α to F and then "closing" the resulting set with respect to addition and multiplication, which amounts to adjoining the powers $\alpha^2, \alpha^3, \ldots$ of α and taking linear combinations (with coefficients from F) of these elements. The process terminates when a power of α is a linear combination of lower powers of α which amounts to knowing the minimal polynomial for α. The previous discussion shows a similar process gives the field $F(\alpha, \beta)$ generated by two elements, and by recursion, the field generated by any finite number of algebraic elements. This shows in particular that "closing" with respect to addition and multiplication also closes with respect to division for algebraic elements (cf. Example 5 following Corollary 5 above). If the elements are not algebraic, one must also "close" with respect to inverses. The difficulty in this procedure is determining the degrees of the *relative* extensions – for example the degree d for $F(\alpha, \beta)$ over $F(\alpha)$ above, for which one has only an a priori upper bound (the degree of β over F).

This is the analogue of "closing" a set of elements in a group G to determine the subgroup they generate.

Examples

(1) The extension $\mathbb{Q}(\sqrt[6]{2}, \sqrt{2})$ is simply the extension $\mathbb{Q}(\sqrt[6]{2})$ since $\sqrt{2}$ is already an element of this field. Put another way, the degree d of $\sqrt{2}$ over $\mathbb{Q}(\sqrt[6]{2})$ is 1, which is strictly smaller than the degree of $\sqrt{2}$ over $\mathbb{Q}$. We shall later have less obvious examples where this occurs.

(2) Consider the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ generated over $\mathbb{Q}$ by $\sqrt{2}$ and $\sqrt{3}$. Since $\sqrt{3}$ is of degree 2 over $\mathbb{Q}$ the degree of the extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}(\sqrt{2})$ is at most 2 and is precisely 2 if and only if $x^2 - 3$ is irreducible over $\mathbb{Q}(\sqrt{2})$. Since this polynomial is of degree 2, it is reducible only if it has a root, i.e. if and only if $\sqrt{3} \in \mathbb{Q}(\sqrt{2})$. Suppose $\sqrt{3} = a + b\sqrt{2}$ with $a, b \in \mathbb{Q}$. Squaring this we obtain $3 = (a^2 + 2b^2) + 2ab\sqrt{2}$. If $ab \neq 0$, then we can solve this equation for $\sqrt{2}$ in terms of a and b which implies that $\sqrt{2}$ is rational, which it is not. If $b = 0$, then we would have that $\sqrt{3} = a$ is rational, a contradiction. Finally, if $a = 0$, we have $\sqrt{3} = b\sqrt{2}$ and multiplying both sides by $\sqrt{2}$ we see that $\sqrt{6}$ would be rational, again a contradiction. This shows $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$, proving

$$[\mathbb{Q}(\sqrt{2}, \sqrt{3}) : \mathbb{Q}] = 4.$$

Elements in this field (by "closing " $1, \sqrt{2}, \sqrt{3}$) include $1, \sqrt{2}, \sqrt{3}, \sqrt{6}$ and by the computations above, these form a basis for this field:

$$\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \{a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mid a, b, c, d \in \mathbb{Q}\}.$$

We can now characterize the finite extensions of a field F:

Theorem 17. The extension K/F is finite if and only if K is generated by a finite number of algebraic elements over F. More precisely, a field generated over F by a finite number of algebraic elements of degrees $n_1, n_2, ..., n_k$ is algebraic of degree $\leq n_1 n_2 ... n_k$.

Proof: If K/F is finite of degree n, let $\alpha_1, \alpha_2, ..., \alpha_n$ be a basis for K as a vector space over F. By Corollary 15 $[F(\alpha_i) : F]$ divides $[K : F] = n$ for $i = 1, 2, ..., n$, so that Proposition 12 implies each α_i is algebraic over F. Since K is obviously generated over F by $\alpha_1, \alpha_2, ..., \alpha_n$, we see that K is generated by a finite number of algebraic elements over F. The converse was proved above. The second statement of the theorem is immediate from Corollary 13 and the computation above.

The first example above shows that the inequality for the degree of the extension given in the theorem may be strict. We remark that information helpful in the determination of this degree can often be obtained by determining subfields and then applying Corollary 15.

Corollary 18. Suppose α and β are algebraic over F. Then $\alpha \pm \beta$, $\alpha\beta$, α/β (for $\beta \neq 0$), (in particular α^{-1} for $\alpha \neq 0$) are all algebraic.

Proof: All of these elements lie in the extension $F(\alpha, \beta)$, which is finite over F by the theorem, hence they are algebraic by Corollary 13.

Corollary 19. Let L/F be an arbitrary extension. Then the collection of elements of L which are algebraic over F form a subfield K of L.

Proof: This is immediate from the previous corollary.

Examples

(1) Consider the extension $\mathbb{C}/\mathbb{Q}$ and let $\overline{\mathbb{Q}}$ denote the subfield of all elements in $\mathbb{C}$ which are algebraic over $\mathbb{Q}$. In particular, the elements $\sqrt[n]{2}$ (the positive n^{th} roots of 2 in $\mathbb{R}$) are all elements of $\overline{\mathbb{Q}}$, so that n divides $[\overline{\mathbb{Q}} : \mathbb{Q}]$ for all integers $n > 1$. Hence $\overline{\mathbb{Q}}$ is an *infinite* algebraic extension of $\mathbb{Q}$.

(2) Consider the field $\overline{\mathbb{Q}} \cap \mathbb{R}$, the subfield of $\mathbb{R}$ consisting of elements algebraic over $\mathbb{Q}$. The field $\mathbb{Q}$ is *countable*. The number of polynomials in $\mathbb{Q}[x]$ of any given degree n is therefore also countable (since such a polynomial is determined by specifying $n + 1$ coefficients from $\mathbb{Q}$). Since these polynomials have at most n roots in $\mathbb{R}$, the number of algebraic elements of $\mathbb{R}$ of degree n is countable. Finally, the collection of all algebraic elements in $\mathbb{R}$ is the countable union (indexed by n) of countable sets, hence is countable. Since $\mathbb{R}$ is uncountable, it follows that there exist (in fact many) elements of $\mathbb{R}$ which are not algebraic, i.e. are transcendental, over $\mathbb{Q}$. In particular the subfield $\overline{\mathbb{Q}} \cap \mathbb{R}$ of algebraic elements of $\mathbb{R}$ is a *proper* subfield of $\mathbb{R}$, so also $\overline{\mathbb{Q}}$ is a proper subfield of $\mathbb{C}$.

It is extremely difficult in general to prove that a given real number is not algebraic. For example, it is known (these are theorems) that $\pi = 3.14159...$ and $e = 2.71828...$ are transcendental elements of $\mathbb{R}$. Even the proofs that these elements are not *rational* are not too easy.

Theorem 20. If K is algebraic over F and L is algebraic over K, then L is algebraic over F.

Proof: Let α be any element of L. Then α is algebraic over K, so α satisfies some polynomial equation

$$a_n\alpha^n + a_{n-1}\alpha^{n-1} + \ldots + a_1\alpha + a_0 = 0$$

where the coefficients $a_0, a_1, \ldots, a_n$ are elements of K. Consider the field $F(\alpha, a_0, a_1, \ldots, a_n)$ generated over F by α and the coefficients of this polynomial. Since K/F is algebraic, the elements $a_0, a_1, \ldots, a_n$ are algebraic over F, so the extension $F(a_0, a_1, \ldots, a_n)/F$ is finite by Theorem 17. By the equation above, we see that α generates an extension of this field of degree at most n, since its minimal polynomial over this field is a divisor of the polynomial above. Therefore

$$[F(\alpha, a_0, a_1, \ldots, a_n) : F] = [F(\alpha, a_0, a_1, \ldots, a_n) : F(a_0, a_1, \ldots, a_n)][F(a_0, a_1, \ldots, a_n) : F$$

is also finite and $F(\alpha, a_0, a_1, \ldots, a_n)/F$ is an algebraic extension. In particular the element α is algebraic over F, which proves that L is algebraic over F.

The subfield $F(\alpha_1, \alpha_2, \ldots, \alpha_k)$ generated by a finite set of elements $\alpha_1, \alpha_2, \ldots, \alpha_k$ of a field K contains each of the fields $F(\alpha_i)$, $i = 1, 2, \ldots, k$. By the definitions, it is also the smallest subfield of K containing these fields.

Definition. Let K_1 and K_2 be two subfields of a field K. Then the *composite field* of K_1 and K_2, denoted K_1K_2, is the smallest subfield of K containing both K_1 and K_2. Similarly, the composite of any collection of subfields of K is the smallest subfield containing all the subfields.

Note that the composite K_1K_2 can also be described as the intersection of all the subfields of K containing both K_1 and K_2 and similarly for the composite of more than two fields, analogous to the subgroup generated by a subset of a group (cf. Section 2.4).

Example

The composite of the two fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt[3]{2})$ is the field $\mathbb{Q}(\sqrt[6]{2})$. This is because this field contains both of these subfields ($(\sqrt[6]{2})^3 = \sqrt{2}$ and $(\sqrt[6]{2})^2 = \sqrt[3]{2}$) and conversely, any field containing both $\sqrt{2}$ and $\sqrt[3]{2}$ contains their quotient, which is $\sqrt[6]{2}$.

Suppose now that K_1 and K_2 are finite extensions of F in K. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be an F-basis for K_1 and let $\beta_1, \beta_2, \ldots, \beta_m$ be an F-basis for K_2 (so that $[K_1 : F] = n$ and $[K_2 : F] = m$). Then it is clear that these give generators for the composite K_1K_2 over F:

$$K_1K_2 = F(\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_m).$$

Since $\alpha_1, \alpha_2, \ldots, \alpha_n$ is an F-basis for K_1 any power ${\alpha_i}^k$ of one of the α's is a *linear* combination with coefficients in F of the α's and a similar statement holds for the β's.

It follows that the collection of linear combinations

$$\sum_{\substack{i=1,2,\ldots,n \\ j=1,2,\ldots,m}} a_{ij}\alpha_i\beta_j$$

with coefficients in F is *closed* under multiplication and addition since in a product of two such elements any higher powers of the α's and β's can be replaced by linear expressions. Hence, the elements $\alpha_i\beta_j$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$ *span* the composite extension K_1K_2 over F. In particular, $[K_1K_2 : F] \leq mn$. We summarize this as:

Proposition 21. Let K_1 and K_2 be two finite extensions of a field F contained in K. Then

$$[K_1K_2 : F] \leq [K_1 : F][K_2 : F]$$

with equality if and only if an F-basis for one of the fields remains linearly independent over the other field. If $\alpha_1, \alpha_2, \ldots, \alpha_n$ and $\beta_1, \beta_2, \ldots, \beta_m$ are bases for K_1 and K_2 over F, respectively, then the elements $\alpha_i\beta_j$ for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$ span K_1K_2 over F.

Proof: From $K_1K_2 = F(\alpha_1, \alpha_2, \ldots, \alpha_n, \beta_1, \beta_2, \ldots, \beta_m) = K_1(\beta_1, \beta_2, \ldots, \beta_m)$, we see as above that $\beta_1, \beta_2, \ldots, \beta_m$ span K_1K_2 over K_1. Hence $[K_1K_2 : K_1] \leq m = [K_2 : F]$ with equality if and only if these elements are linearly independent over K_1. Since $[K_1K_2 : F] = [K_1K_2 : K_1][K_1 : F]$ this proves the proposition.

By the proposition (and its proof), we have the following diagram:

$$\begin{tikzcd}
 & K_1K_2 \arrow[dl, dash, "\leq m"'] \arrow[dr, dash, "\leq n"] & \\
K_1 \arrow[dr, dash, "n"'] & & K_2 \arrow[dl, dash, "m"] \\
 & F &
\end{tikzcd}$$

We shall have examples shortly where the inequality in the proposition is strict. One useful situation where one can be certain of equality is the following:

Corollary 22. Suppose that $[K_1 : F] = n$, $[K_2 : F] = m$ in Proposition 21, where n and m are relatively prime: $(n, m) = 1$. Then $[K_1K_2 : F] = [K_1 : F][K_2 : F] = nm$.

Proof: In general the extension degree $[K_1K_2 : F]$ is divisible by both n and m since K_1 and K_2 are subfields of K_1K_2, hence is divisible by their least common multiple. In this case, since $(n, m) = 1$, this means $[K_1K_2 : F]$ is divisible by nm, which together with the inequality $[K_1K_2 : F] \leq nm$ of the proposition proves the corollary.

Example

The composite of the two fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt[3]{2})$ is of degree 6 over $\mathbb{Q}$, which we determined earlier by actually computing the composite $\mathbb{Q}(\sqrt[6]{2})$.

EXERCISES

1. Let $\mathbb{F}$ be a finite field of characteristic p. Prove that $|\mathbb{F}| = p^n$ for some positive integer n.
2. Let $g(x) = x^2 + x - 1$ and let $h(x) = x^3 - x + 1$. Obtain fields of 4,8,9 and 27 elements by adjoining a root of $f(x)$ to the field F where $f(x) = g(x)$ or $h(x)$ and $F = \mathbb{F}_2$ or $\mathbb{F}_3$. Write down the multiplication tables for the fields with 4 and 9 elements and show that the nonzero elements form a cyclic group.
3. Determine the minimal polynomial over $\mathbb{Q}$ for the element $1 + i$.
4. Determine the degree over $\mathbb{Q}$ of $2 + \sqrt{3}$ and of $1 + \sqrt[3]{2} + \sqrt[3]{4}$.
5. Let $F = \mathbb{Q}(i)$. Prove that $x^3 - 2$ and $x^3 - 3$ are irreducible over F.
6. Prove directly from the definitions that the field $F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ is the composite of the fields $F(\alpha_1), F(\alpha_2), \ldots, F(\alpha_n)$.
7. Prove that $\mathbb{Q}(\sqrt{2} + \sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ [one inclusion is obvious, for the other consider $(\sqrt{2} + \sqrt{3})^2$, etc.]. Conclude that $[\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}] = 4$. Find an irreducible polynomial satisfied by $\sqrt{2} + \sqrt{3}$.
8. Let F be a field of characteristic $\neq 2$. Let D_1 and D_2 be elements of F, neither of which is a square in F. Prove that $F(\sqrt{D_1}, \sqrt{D_2})$ is of degree 4 over F if $D_1 D_2$ is not a square in F and is of degree 2 over F otherwise. When $F(\sqrt{D_1}, \sqrt{D_2})$ is of degree 4 over F the field is called a *biquadratic extension of F*.
9. Let F be a field of characteristic $\neq 2$. Let a, b be elements of the field F with b not a square in F. Prove that a necessary and sufficient condition for $\sqrt{a + \sqrt{b}} = \sqrt{m} + \sqrt{n}$ for some m and n in F is that $a^2 - b$ is a square in F. Use this to determine when the field $\mathbb{Q}(\sqrt{a + \sqrt{b}})$ $(a, b \in \mathbb{Q})$ is biquadratic over $\mathbb{Q}$.
10. Determine the degree of the extension $\mathbb{Q}(\sqrt{3 + 2\sqrt{2}})$ over $\mathbb{Q}$.
11. **(a)** Let $\sqrt{3 + 4i}$ denote the square root of the complex number $3 + 4i$ that lies in the first quadrant and let $\sqrt{3 - 4i}$ denote the square root of $3 - 4i$ that lies in the fourth quadrant. Prove that $[\mathbb{Q}(\sqrt{3 + 4i} + \sqrt{3 - 4i}) : \mathbb{Q}] = 1$.

    (b) Determine the degree of the extension $\mathbb{Q}(\sqrt{1 + \sqrt{-3}} + \sqrt{1 - \sqrt{-3}})$ over $\mathbb{Q}$.
12. Suppose the degree of the extension K/F is a prime p. Show that any subfield E of K containing F is either K or F.
13. Suppose $F = \mathbb{Q}(\alpha_1, \alpha_2, \ldots, \alpha_n)$ where $\alpha_i^2 \in \mathbb{Q}$ for $i = 1, 2, \ldots, n$. Prove that $\sqrt[3]{2} \notin F$.
14. Prove that if $[F(\alpha) : F]$ is odd then $F(\alpha) = F(\alpha^2)$.
15. A field F is said to be *formally real* if -1 is not expressible as a sum of squares in F. Let F be a formally real field. Let $f(x) \in F[x]$ be an irreducible polynomial of odd degree and let α be a root of $f(x)$. Prove that $F(\alpha)$ is also formally real.
16. Let K/F be an algebraic extension and let R be a *ring* contained in K and containing F. Show that R is a subfield of K containing F.
17. Let $f(x)$ be an irreducible polynomial of degree n over a field F. Let $g(x)$ be any polynomial in $F[x]$. Prove that every irreducible factor of the composite polynomial $f(g(x))$ has degree divisible by n.

18. Let F be a field and let $F(x)$ be the field of rational functions in x with coefficients from F. Let $t \in F(x)$ be the rational function $\dfrac{P(x)}{Q(x)}$ with relatively prime polynomials $P(x), Q(x) \in F[x]$, with $Q(x) \neq 0$. Then $F(x)$ is an extension of $F(t)$ and to compute its degree it is necessary to compute the minimal polynomial with coefficients in $F(t)$ satisfied by x.
   (a) Show that the polynomial $P(X) - tQ(X)$ in the variable X and coefficients in $F(t)$ is irreducible over $F(t)$ and has x as a root. [By Gauss' Lemma this polynomial is irreducible in $(F(t))[X]$ if and only if it is irreducible in $(F[t])[X]$. Then note that $(F[t])[X] = (F[X])[t]$.]
   (b) Show that the degree of $P(X) - tQ(X)$ as a polynomial in X with coefficients in $F(t)$ is the maximum of the degrees of $P(x)$ and $Q(x)$.
   (c) Show that $[F(x) : F(t)] = [F(x) : F(\dfrac{P(x)}{Q(x)})] = \max(\deg P(x), \deg Q(x))$.

19. Let K be an extension of F of degree n.
   (a) For any $\alpha \in K$ prove that α acting by left multiplication on K is an F-linear transformation of K.
   (b) Prove that K is isomorphic to a subfield of the ring of $n \times n$ matrices over F, so the ring of $n \times n$ matrices over F contains an isomorphic copy of *every* extension of F of degree $\leq n$.

20. Show that if the matrix of the linear transformation "multiplication by α" considered in the previous exercise is A then α is a root of the characteristic polynomial for A. This gives an effective procedure for determining an equation of degree n satisfied by an element α in an extension of F of degree n. Use this procedure to obtain the monic polynomial of degree 3 satisfied by $\sqrt[3]{2}$ and by $1 + \sqrt[3]{2} + \sqrt[3]{4}$.

21. Let $K = \mathbb{Q}(\sqrt{D})$ for some squarefree integer D. Let $\alpha = a + b\sqrt{D}$ be an element of K. Use the basis $1, \sqrt{D}$ for K as a vector space over $\mathbb{Q}$ and show that the matrix of the linear transformation "multiplication by α" on K considered in the previous exercises has the matrix $\begin{pmatrix} a & bD \\ b & a \end{pmatrix}$. Prove directly that the map $a + b\sqrt{D} \mapsto \begin{pmatrix} a & bD \\ b & a \end{pmatrix}$ is an isomorphism of the field K with a subfield of the ring of 2×2 matrices with coefficients in $\mathbb{Q}$.

13.3 CLASSICAL STRAIGHTEDGE AND COMPASS CONSTRUCTIONS

As a simple application of the results we have obtained on algebraic extensions, and in particular on the multiplicativity of extension degrees, we can answer (in the negative) the following geometric problems posed by the Greeks:

I. (Doubling the Cube) Is it possible using only straightedge and compass to construct a cube with precisely twice the volume of a given cube?

II. (Trisecting an Angle) Is it possible using only straightedge and compass to trisect any given angle θ?

III. (Squaring the Circle) Is it possible using only straightedge and compass to construct a square whose area is precisely the area of a given circle?

To answer these questions we must translate the construction of lengths by compass and straightedge into algebraic terms. Let 1 denote a fixed given unit distance.

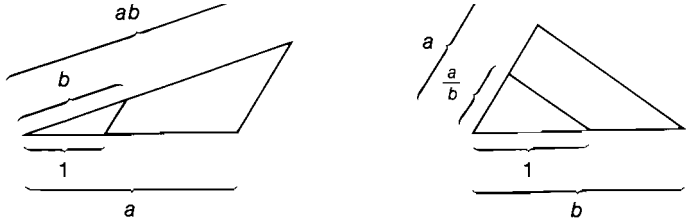


Fig. 1

Then any distance is determined by its length $a \in \mathbb{R}$, which allows us to view geometric distances as elements of the real numbers $\mathbb{R}$. Using the given unit distance 1 to define the scale on the axes, we can then construct the usual Cartesian plane $\mathbb{R}^2$ and view all of our constructions as occurring in $\mathbb{R}^2$. A point $(x, y) \in \mathbb{R}^2$ is then constructible starting with the given distance 1 if and only if its coordinates x and y are constructible elements of $\mathbb{R}$. The problems above then amount to determining whether particular lengths in $\mathbb{R}$ can be obtained by compass and straightedge constructions from a fixed unit distance. The collection of such real numbers together with their negatives will be called the *constructible* elements of $\mathbb{R}$, and we shall not distinguish between the lengths which are constructible and the real numbers which are constructible.

Each straightedge and compass construction consists of a series of operations of the following four types: (1) connecting two given points by a straight line, (2) finding a point of intersection of two straight lines, (3) drawing a circle with given radius and center, and (4) finding the point(s) of intersection of a straight line and a circle or the intersection of two circles.

It is an elementary fact from geometry that if two lengths a and b are given one may construct using straightedge and compass the lengths $a \pm b$, ab and a/b (the first two are clear and the latter two are given by the construction of parallel lines (Fig. 1)).

It is also an elementary geometry construction to construct $\sqrt{a}$ if a is given: construct the circle with diameter $1 + a$ and erect the perpendicular to the diameter as indicated in Fig. 2. Then $\sqrt{a}$ is the length of this perpendicular.

It follows that straightedge and compass constructions give all the algebraic operations of addition, subtraction, multiplication and division (by nonzero elements) in the reals so the collection of constructible elements is a *subfield* of $\mathbb{R}$. One can also take square roots of constructible elements. We shall now see that these are essentially the only operations possible.

From the given length 1 it is possible to construct by these operations all the rational numbers $\mathbb{Q}$. Hence we may construct all of the points $(x, y) \in \mathbb{R}^2$ whose

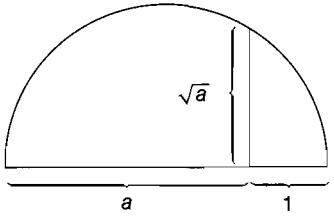


Fig. 2

coordinates are rational. We may construct additional elements of $\mathbb{R}$ by taking square roots, so the collection of elements constructible from 1 of $\mathbb{R}$ form a field strictly larger than $\mathbb{Q}$.

The usual formula ("two point form") for the straight line connecting two points with coordinates in some field F gives an equation for the line of the form $ax+by-c = 0$ with $a, b, c \in F$. Solving two such equations simultaneously to determine the point of intersection of two such lines gives solutions also in F. It follows that if the coordinates of two points lie in the field F then straightedge constructions alone will not produce additional points whose coordinates are not also in F.

A compass construction (type (3) or (4) above) defines points obtained by the intersection of a circle with either a straight line or another circle. A circle with center (h, k) and radius r has equation

$$(x-h)^2 + (y-k)^2 = r^2$$

so when we consider the effect of compass constructions on elements of a field F we are considering simultaneous solutions of such an equation with a linear equation $ax + by - c = 0$ where $a, b, c, h, k, r \in F$, or the simultaneous solutions of two quadric equations.

In the case of a linear equation and the equation for the circle, solving for y, say, in the linear equation and substituting gives a *quadratic* equation for x (and y is given linearly in terms of x). Hence the coordinates of the point of intersection are at worst in a *quadratic extension* of F.

In the case of the intersection of two circles, say

$$\begin{aligned} &(x-h)^2 + (y-k)^2 = r^2 \\ \text{and} \qquad &(x-h')^2 + (y-k')^2 = r'^2, \end{aligned}$$

subtraction of the second equation from the first shows that we have the same intersection by considering the two equations

$$\begin{aligned} &(x-h)^2 + (y-k)^2 = r^2 \\ \text{and} \qquad &2(h'-h)x + 2(k'-k)y = r^2 - h^2 - k^2 - r'^2 + h'^2 + k'^2, \end{aligned}$$

which is the intersection of a circle and a straight line (the straight line connecting the two points of intersection, in fact) of the type just considered.

It follows that if a collection of constructible elements is given, then one can construct all the elements in the subfield F of $\mathbb{R}$ generated by these elements and that any straightedge and compass operation on elements of F produces elements in at worst a *quadratic* extension of F. Since quadratic extensions have degree 2 and extension degrees are multiplicative, it follows that if $\alpha \in \mathbb{R}$ is obtained from elements in a field F by a (finite) series of straightedge and compass operations then α is an element of an extension K of F of degree a power of 2: $[K : F] = 2^m$ for some m. Since $[F(\alpha) : F]$ divides this extension degree, it must also be a power of 2.

Proposition 23. If the element $\alpha \in \mathbb{R}$ is obtained from a field $F \subset \mathbb{R}$ by a series of compass and straightedge constructions then $[F(\alpha) : F] = 2^k$ for some integer $k \geq 0$.

Theorem 24. None of the classical Greek problems : (I) Doubling the Cube, (II) Trisecting an Angle, and (III) Squaring the Circle , is possible.

Proof: (I) Doubling the cube amounts to constructing $\sqrt[3]{2}$ in the reals starting with the unit 1. Since $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ is not a power of 2, this is impossible.

(II) If an angle θ can be constructed, then determining the point at distance 1 from the origin and angle θ from the positive x axis in $\mathbb{R}^2$ shows that $\cos\theta$ (the x-coordinate of this point) can be constructed (so then $\sin\theta$ can also be constructed). Conversely if $\cos\theta$, then $\sin\theta$, can be constructed, the point with those coordinates gives the angle θ.

The problem of trisecting the angle θ is then equivalent to the problem: given $\cos\theta$ construct $\cos\theta/3$.

To see that this is not always possible (it is certainly occasionally possible, for example for $\theta = 180°$), consider $\theta = 60°$. Then $\cos\theta = \frac{1}{2}$. By the triple angle formula for cosines:

$$\cos\theta = 4\cos^3\theta/3 - 3\cos\theta/3,$$

substituting $\theta = 60°$, we see that $\beta = \cos 20°$ satisfies the equation

$$4\beta^3 - 3\beta - 1/2 = 0$$

or $8(\beta)^3 - 6\beta - 1 = 0$. This can be written $(2\beta)^3 - 3(2\beta) - 1 = 0$. Let $\alpha = 2\beta$. Then α is a real number between 0 and 2 satisfying the equation

$$\alpha^3 - 3\alpha - 1 = 0.$$

But we considered this equation in the last section and determined $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 3$, and as before we see that α is not constructible.

(III) Squaring the circle is equivalent to determining whether the real number $\pi = 3.14159...$ is constructible. As mentioned previously, it is a difficult problem even to prove that this number is not rational. It is in fact transcendental (which we shall assume without proof), so that $[\mathbb{Q}(\pi) : \mathbb{Q}]$ is not even finite, much less a power of 2, showing the impossibility of squaring the circle by straightedge and compass.

Remark: The proof above shows that $\cos 20°$ and $\sin 20°$ cannot be constructed. The question arises as to which integer angles (measured in degrees) are constructible? The angles 1° and 2° are not constructible, since otherwise the addition formulae for sines and cosines would give the constructibility for 20°. On the other hand, elementary geometric constructions (of the regular 5-gon for an angle of 72° and the equilateral triangle for an angle of 60°) together with the addition formulae and the half-angle formulae show that $\cos 3°$ and $\sin 3°$ are constructible. It follows from this that the trigonometric functions of an integer degree angle are constructible precisely when the angle is a multiple of 3°. Explicitly:

$$\cos 3° = \frac{1}{8}(\sqrt{3}+1)\sqrt{5+\sqrt{5}} + \frac{1}{16}(\sqrt{6}-\sqrt{2})(\sqrt{5}-1)$$
$$\sin 3° = \frac{1}{16}(\sqrt{6}+\sqrt{2})(\sqrt{5}-1) - \frac{1}{8}(\sqrt{3}-1)\sqrt{5+\sqrt{5}}$$

showing that these are obtained from $\mathbb{Q}$ by successive extractions of square roots and field operations.

After discussing the cyclotomic fields in Section 14.5 we shall consider another classical geometric question: which regular n-gons can be constructed by straightedge and compass? (cf. Proposition 14.29).

We have been careful here to consider constructions using a *straightedge* rather than a *ruler*, the distinction being that a ruler has marks on it. If one uses a ruler, it is possible to construct many additional algebraic elements.

For example, suppose θ is a given angle and the unit distance 1 is marked on the ruler. Draw a circle of radius 1 with central angle θ as shown in Fig. 3 and then slide the ruler until the distance between points A and B on the circle is 1. Then some elementary geometry shows that (cf. the exercises) the angle α indicated is $\theta/3$, i.e. this construction (due to Archimedes) trisects θ. In particular, the second classical problem in Theorem 24 (Trisecting an Angle) can be solved with *ruler* and compass.

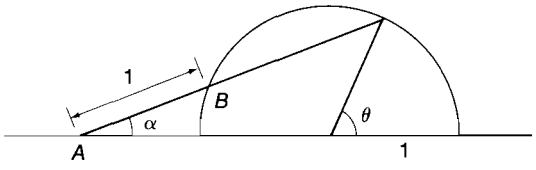


Fig. 3

The first of the classical problems in Theorem 24 (Duplication of the Cube), which amounts to the construction of $\sqrt[3]{2}$, can also be solved with ruler and compass. The following gives a construction for $k^{1/3}$ for any given positive real k which is less than 1. This construction was shown to us by J. Conway.

Drawing a circle of radius 1 and using the point $A = (k, 0)$ as center, construct the point $B = (0, \sqrt{1-k^2})$. Dividing this distance by 3, construct the point $(0, -\frac{1}{3}\sqrt{1-k^2})$ and draw the line connecting this point with A. Slide the ruler with marked unit length 1 so that it passes through the point B and so that the distance from the intersection point C to the intersection point D with the x–axis is of length 1, as indicated in Fig. 4.

Then the distance between A and D is $2k^{1/3}$ and the distance between B and C is $2k^{2/3}$ (cf. the exercises).

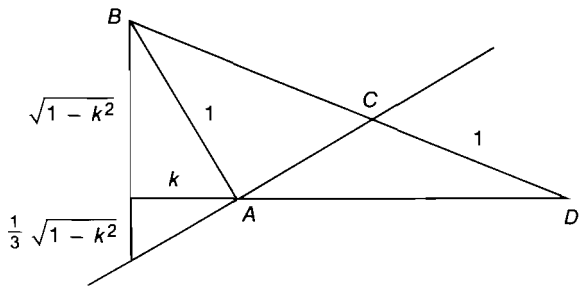


Fig. 4

EXERCISES

1. Prove that it is impossible to construct the regular 9–gon.
2. Prove that Archimedes' construction actually trisects the angle θ. [Note the isosceles triangles in Fig. 5 to prove that $\beta = \gamma = 2\alpha$.]

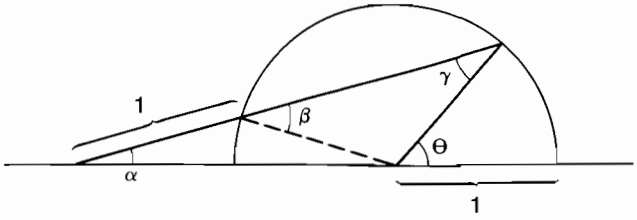


Fig. 5

3. Prove that Conway's construction indicated in the text actually constructs $2k^{1/3}$ and $2k^{2/3}$. [One method: let (x, y) be the coordinates of the point C, a the distance from B to C and b the distance from A to D; use similar triangles to prove (a) $\dfrac{y}{1} = \dfrac{\sqrt{1-k^2}}{1+a}$, (b) $\dfrac{x}{a} = \dfrac{b+k}{1+a}$, (c) $\dfrac{y}{x-k} = \dfrac{\sqrt{1-k^2}}{3k}$, and also show that (d) $(1-k^2) + (b+k)^2 = (1+a)^2$; solve these equations for a and b.]
4. The construction of the regular 7–gon amounts to the constructibility of $\cos(2\pi/7)$. We shall see later (Section 14.5 and Exercise 2 of Section 14.7) that $\alpha = 2\cos(2\pi/7)$ satisfies the equation $x^3 + x^2 - 2x - 1 = 0$. Use this to prove that the regular 7–gon is not constructible by straightedge and compass.
5. Use the fact that $\alpha = 2\cos(2\pi/5)$ satisfies the equation $x^2 + x - 1 = 0$ to conclude that the regular 5–gon is constructible by straightedge and compass.

13.4 SPLITTING FIELDS AND ALGEBRAIC CLOSURES

Let F be a field.

If $f(x)$ is any polynomial in $F[x]$ then we have seen in Section 2 that there exists a field K which can (by identifying F with an isomorphic copy of F) be considered an extension of F in which $f(x)$ has a root α. This is equivalent to the statement that $f(x)$ has a linear factor $x - \alpha$ in $K[x]$ (this is Proposition 10 of Chapter 9).

Definition. The extension field K of F is called a *splitting field* for the polynomial $f(x) \in F[x]$ if $f(x)$ factors completely into linear factors (or *splits completely*) in $K[x]$ and $f(x)$ does not factor completely into linear factors over any proper subfield of K containing F.

If $f(x)$ is of degree n, then $f(x)$ has at most n roots in F (Proposition 17 of Chapter 9) and has precisely n roots (counting multiplicities) in F if and only if $f(x)$ splits completely in $F[x]$.

Theorem 25. For any field F, if $f(x) \in F[x]$ then there exists an extension K of F which is a splitting field for $f(x)$.

Proof: We first show that there is an extension E of F over which $f(x)$ splits completely into linear factors by induction on the degree n of $f(x)$. If $n = 1$, then take $E = F$. Suppose now that $n > 1$. If the irreducible factors of $f(x)$ over F are all of degree 1, then F is the splitting field for $f(x)$ and we may take $E = F$. Otherwise, at least one of the irreducible factors, say $p(x)$ of $f(x)$ in $F[x]$ is of degree at least 2. By Theorem 3 there is an extension E_1 of F containing a root α of $p(x)$. Over E_1 the polynomial $f(x)$ has the linear factor $x - \alpha$. The degree of the remaining factor $f_1(x)$ of $f(x)$ is $n - 1$, so by induction there is an extension E of E_1 containing all the roots of $f_1(x)$. Since $\alpha \in E$, E is an extension of F containing all the roots of $f(x)$. Now let K be the intersection of all the subfields of E containing F which also contain all the roots of $f(x)$. Then K is a field which is a splitting field for $f(x)$.

We shall see shortly that any two splitting fields for $f(x)$ are isomorphic (which extends Theorem 8), so (by abuse) we frequently refer to *the* splitting field of a polynomial.

Definition. If K is an algebraic extension of F which is the splitting field over F for a collection of polynomials $f(x) \in F[x]$ then K is called a *normal* extension of F.

We shall generally use the term "splitting field" rather than "normal extension" (cf. also Section 14.9).

Examples

(1) The splitting field for $x^2 - 2$ over $\mathbb{Q}$ is just $\mathbb{Q}(\sqrt{2})$, since the two roots are $\pm\sqrt{2}$ and $-\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.

(2) The splitting field for $(x^2 - 2)(x^2 - 3)$ is the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ generated over $\mathbb{Q}$ by $\sqrt{2}$ and $\sqrt{3}$ since the roots of the polynomial are $\pm\sqrt{2}, \pm\sqrt{3}$. We have already seen that this is an extension of degree 4 over $\mathbb{Q}$ and we have the following diagram of known subfields:

$$\begin{tikzcd}
 & \mathbb{Q}(\sqrt{2}, \sqrt{3}) \arrow[dl, dash, "2"'] \arrow[d, dash, "2"] \arrow[dr, dash, "2"] & \\
\mathbb{Q}(\sqrt{2}) \arrow[dr, dash, "2"'] & \mathbb{Q}(\sqrt{6}) \arrow[d, dash, "2"] & \mathbb{Q}(\sqrt{3}) \arrow[dl, dash, "2"] \\
 & \mathbb{Q} &
\end{tikzcd}$$

(3) The splitting field of $x^3 - 2$ over $\mathbb{Q}$ is not just $\mathbb{Q}(\sqrt[3]{2})$ since as previously noted the three roots of this polynomial in $\mathbb{C}$ are

$$\sqrt[3]{2}, \quad \sqrt[3]{2}\left(\frac{-1 + i\sqrt{3}}{2}\right), \quad \sqrt[3]{2}\left(\frac{-1 - i\sqrt{3}}{2}\right)$$

and the latter two roots are not elements of $\mathbb{Q}(\sqrt[3]{2})$, since the elements of this field are of the form $a + b\sqrt[3]{2} + c\sqrt[3]{4}$ with rational a, b, c and all such numbers are real.

The splitting field K of this polynomial is obtained by adjoining all three of these roots to $\mathbb{Q}$. Note that since K contains the first two roots above, then it contains their quotient $\dfrac{-1+\sqrt{-3}}{2}$ hence K contains the element $\sqrt{-3}$. On the other hand, any field containing $\sqrt[3]{2}$ and $\sqrt{-3}$ contains all three of the roots above. It follows that

$$K = \mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})$$

is the splitting field of $x^3 - 2$ over $\mathbb{Q}$. Since $\sqrt{-3}$ satisfies the equation $x^2 + 3 = 0$, the degree of this extension over $\mathbb{Q}(\sqrt[3]{2})$ is at most 2, hence must be 2 since we observed above that $\mathbb{Q}(\sqrt[3]{2})$ is not the splitting field. It follows that

$$[\mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) : \mathbb{Q}] = 6.$$

Note that we could have proceeded slightly differently at the end by noting that $\mathbb{Q}(\sqrt{-3})$ is a subfield of K, so that the index $[\mathbb{Q}(\sqrt{-3}) : \mathbb{Q}] = 2$ divides $[K : \mathbb{Q}]$. Since this extension degree is also divisible by 3 (because $\mathbb{Q}(\sqrt[3]{2}) \subset K$), the degree is divisible by 6, hence must be 6.

This gives us the diagram of known subfields:

$$\begin{tikzcd}
 & \mathbb{Q}(\sqrt[3]{2}, \sqrt{-3}) \arrow[dl, dash] \arrow[dr, dash] \arrow[drr, dash] \arrow[drrr, dash] & & & \\
\mathbb{Q}(\sqrt{-3}) \arrow[dr, dash] & & \mathbb{Q}(\theta_1) \arrow[dl, dash] & \mathbb{Q}(\theta_2) \arrow[dll, dash] & \mathbb{Q}(\theta_3) \arrow[dlll, dash] \\
 & \mathbb{Q} & & &
\end{tikzcd}$$

where

$$\theta_1 = \sqrt[3]{2}, \quad \theta_2 = \sqrt[3]{2}\left(\frac{-1+i\sqrt{3}}{2}\right), \quad \theta_3 = \sqrt[3]{2}\left(\frac{-1-i\sqrt{3}}{2}\right).$$

(4) One must be careful in computing splitting fields. The splitting field for the polynomial $x^4 + 4$ over $\mathbb{Q}$ is smaller than one might at first suspect. In fact this polynomial factors over $\mathbb{Q}$:

$$\begin{aligned} x^4 + 4 &= x^4 + 4x^2 + 4 - 4x^2 = (x^2+2)^2 - 4x^2 \\ &= (x^2 + 2x + 2)(x^2 - 2x + 2) \end{aligned}$$

where these two factors are irreducible (Eisenstein again). Solving for the roots of the two factors by the quadratic formula, we find the four roots

$$\pm 1 \pm i$$

so that the splitting field of this polynomial is just the field $\mathbb{Q}(i)$, an extension of degree 2 of $\mathbb{Q}$.

In general, if $f(x) \in F[x]$ is a polynomial of degree n, then adjoining one root of $f(x)$ to F generates an extension F_1 of degree at most n (and equal to n if and only if $f(x)$ is irreducible). Over F_1 the polynomial $f(x)$ now has at least one linear factor,

so that any other root of $f(x)$ satisfies an equation of degree at most $n-1$ over F_1. Adjoining such a root to F_1 we therefore obtain an extension of degree at most $n-1$ of F_1, etc. Using the multiplicativity of extension degrees, this proves

Proposition 26. A splitting field of a polynomial of degree n over F is of degree at most $n!$ over F.

As the examples above show, the degree of a splitting field may be smaller than $n!$. It will be proved later using Galois Theory that a "general" polynomial of degree n (in a well defined sense) over $\mathbb{Q}$ has a splitting field of degree $n!$, so this may be viewed as the "generic" situation (although most of the interesting examples we shall consider have splitting fields of smaller degree).

Example: (Splitting Field of $x^n - 1$: Cyclotomic Fields)

Consider the splitting field of the polynomial $x^n - 1$ over $\mathbb{Q}$. The roots of this polynomial are called the nth *roots of unity*.

Recall that every nonzero complex number $a + bi \in \mathbb{C}$ can be written uniquely in the form

$$re^{i\theta} = r(\cos\theta + i\sin\theta) \qquad r > 0, \quad 0 \le \theta < 2\pi$$

which is simply representing the point $a + bi$ in the complex plane in terms of polar coordinates: r is the distance of (a, b) from the origin and θ is the angle made with the real axis.

Over $\mathbb{C}$ there are n distinct solutions of the equation $x^n = 1$, namely the elements

$$e^{2\pi ki/n} = \cos\left(\frac{2\pi k}{n}\right) + i\sin\left(\frac{2\pi k}{n}\right)$$

for $k = 0, 1, \ldots, n-1$. These points are given geometrically by n equally spaced points starting with the point $(1,0)$ (corresponding to $k = 0$) on a circle of radius 1 in the complex plane (see Fig. 6). The fact that these are all n^{th} roots of unity is immediate, since

$$(e^{2\pi ki/n})^n = e^{(2\pi ki/n)n} = e^{2\pi ki} = 1.$$

It follows that $\mathbb{C}$ contains a splitting field for $x^n - 1$ and we shall frequently view the splitting field for $x^n - 1$ over $\mathbb{Q}$ as the field generated over $\mathbb{Q}$ in $\mathbb{C}$ by the numbers above.

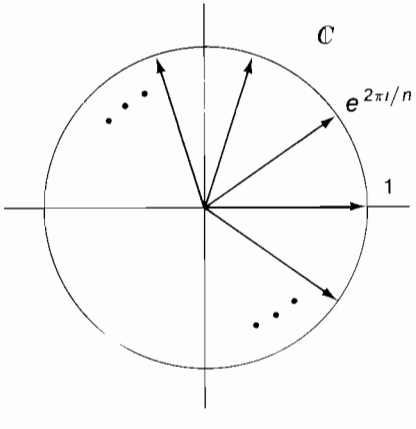


Fig. 6

In *any* abstract splitting field $K/\mathbb{Q}$ for $x^n - 1$ the collection of nth roots of unity form a *group* under multiplication since if $\alpha^n = 1$ and $\beta^n = 1$ then $(\alpha\beta)^n = 1$ so this subset of n elements of $K^\times$ is closed under multiplication. It follows that this is a *cyclic* group (Proposition 18 of Chapter 9) of order n.

Definition. A generator of the cyclic group of all n^{th} roots of unity is called a *primitive* n^{th} root of unity.

Let ζ_n denote a primitive n^{th} root of unity. The other *primitive* n^{th} roots of unity are then the elements ${\zeta_n}^a$ where $1 \leq a < n$ is an integer relatively prime to n, since these are the other generators for a cyclic group of order n. In particular there are precisely $\varphi(n)$ primitive n^{th} roots of unity, where $\varphi(n)$ denotes the Euler φ-function.

Over $\mathbb{C}$ we can see all of this directly by letting

$$\zeta_n = e^{2\pi i/n}$$

(the first n^{th} root of unity counterclockwise from 1). Then all the other roots of unity are powers of ζ_n:

$$e^{2\pi k i/n} = {\zeta_n}^k$$

so that ζ_n is one possible generator for the multiplicative group of nth roots of unity. When we view the roots of unity in $\mathbb{C}$ we shall usually use ζ_n to denote this choice of a primitive n^{th} root of unity. The primitive roots of unity in $\mathbb{C}$ for some small values of n are

$$\begin{aligned}
\zeta_1 &= 1 \\
\zeta_2 &= -1 \\
\zeta_3 &= \frac{-1 + i\sqrt{3}}{2} \\
\zeta_4 &= i \\
\zeta_5 &= \frac{\sqrt{5} - 1}{4} + i\left(\frac{\sqrt{10 + 2\sqrt{5}}}{4}\right) \\
\zeta_6 &= \frac{1 + i\sqrt{3}}{2} \\
\zeta_8 &= \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2}
\end{aligned}$$

(these formulas follow from the elementary geometry of n-gons and in any case can be verified directly by raising them to the appropriate power).

The splitting field of $x^n - 1$ over $\mathbb{Q}$ is the field

$$\mathbb{Q}(\zeta_n).$$

Definition. The field $\mathbb{Q}(\zeta_n)$ is called the *cyclotomic field of* n^{th} *roots of unity*.

Determining the degree of this extension requires some analysis of the minimal polynomial of ζ_n over $\mathbb{Q}$ and will be postponed until later (Section 6). One important special case which we have in fact already considered is when $n = p$ is a *prime*. In this

case, we have the factorization

$$x^p - 1 = (x-1)(x^{p-1} + x^{p-2} + \ldots + x + 1)$$

and since $\zeta_p \neq 1$ it follows that ζ_p is a root of the polynomial

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \ldots + x + 1$$

which we showed was irreducible in Chapter 9.4. It follows that $\Phi_p(x)$ is the minimal polynomial of ζ_p over $\mathbb{Q}$, so that

$$[\mathbb{Q}(\zeta_p) : \mathbb{Q}] = p - 1.$$

We shall see later than in general $[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n)$, where $\varphi(n)$ is the Euler phi-function of n (so that $\varphi(p) = p - 1$).

Example: (Splitting Field $x^p - 2$, p a prime)

Let p be a prime and consider the splitting field of $x^p - 2$. If α is a root of this equation, i.e. $\alpha^p = 2$, then $(\zeta\alpha)^p = 2$ where ζ is any p^{th} root of unity. Hence the solutions of this equation are

$$\zeta\sqrt[p]{2}. \qquad \zeta \text{ a } p^{\text{th}} \text{ root of unity}$$

where as usual the symbol $\sqrt[p]{2}$ denotes the positive real p^{th} root of 2 if we wish to view these elements as complex numbers, and denotes any one solution of $x^p = 2$ if we view these roots abstractly. Since the ratio of the two solutions $\zeta_p\sqrt[p]{2}$ and $\sqrt[p]{2}$ for ζ_p a primitive p^{th} root of unity is just ζ_p, the splitting field of $x^p - 2$ over $\mathbb{Q}$ contains $\mathbb{Q}(\sqrt[p]{2}, \zeta_p)$. On the other hand, all the roots above lie in this field, so that the splitting field is precisely

$$\mathbb{Q}(\sqrt[p]{2}, \zeta_p).$$

This field contains the cyclotomic field of pth roots of unity and is generated over it by $\sqrt[p]{2}$, hence is an extension of degree at most p. It follows that the degree of this extension is $\leq p(p-1)$. Since both $\mathbb{Q}(\sqrt[p]{2})$ and $\mathbb{Q}(\zeta_p)$ are subfields, the degree of the extension over $\mathbb{Q}$ is divisible by p and by $p - 1$. Since these two numbers are relatively prime it follows that the extension degree is divisible by $p(p-1)$ so that we must have

$$[\mathbb{Q}(\sqrt[p]{2}, \zeta_p) : \mathbb{Q}] = p(p-1)$$

(this is Corollary 22). Note in particular that we have proved $x^p - 2$ remains irreducible over $\mathbb{Q}(\zeta_p)$, which is not at all obvious. We have the following diagram of known subfields:

$$\begin{tikzcd}
 & \mathbb{Q}(\sqrt[p]{2}, \zeta_p) \arrow[dl, dash, "p"'] \arrow[dr, dash, "p-1"] & \\
\mathbb{Q}(\zeta_p) \arrow[dr, dash, "p-1"'] & & \mathbb{Q}(\sqrt[p]{2}) \arrow[dl, dash, "p"] \\
 & \mathbb{Q} &
\end{tikzcd}$$

The special case $p = 3$ was Example 3 above, where we simply indicated the 3rd roots of unity explicitly.

We now return to the problem of proving it makes no difference how the splitting field of a polynomial $f(x)$ over a field F is constructed. As in Theorem 8 it is convenient to state the result for an arbitrary isomorphism $\varphi : F \xrightarrow{\sim} F'$ between two fields.

Theorem 27. Let $\varphi : F \xrightarrow{\sim} F'$ be an isomorphism of fields. Let $f(x) \in F[x]$ be a polynomial and let $f'(x) \in F'[x]$ be the polynomial obtained by applying φ to the coefficients of $f(x)$. Let E be a splitting field for $f(x)$ over F and let E' be a splitting field for $f'(x)$ over F'. Then the isomorphism φ extends to an isomorphism $\sigma : E \xrightarrow{\sim} E'$, i.e. σ restricted to F is the isomorphism φ:

$$\begin{array}{ccccc} \sigma: & E & \xrightarrow{\sim} & E' \\ & | & & | \\ \varphi: & F & \xrightarrow{\sim} & F' \end{array}$$

Proof: We shall proceed by induction on the degree n of $f(x)$. As in the discussion before Theorem 8, recall that an isomorphism φ from one field F to another field F' induces a natural isomorphism between the polynomial rings $F[x]$ and $F'[x]$. In particular, if $f(x)$ and $f'(x)$ correspond to one another under this isomorphism then the irreducible factors of $f(x)$ in $F[x]$ correspond to the irreducible factors of $f'(x)$ in $F'[x]$.

If $f(x)$ has all its roots in F then $f(x)$ splits completely in $F[x]$ and $f'(x)$ splits completely in $F'[x]$ (with its linear factors being the images of the linear factors for $f(x)$). Hence $E = F$ and $E' = F'$, and in this case we may take $\sigma = \varphi$. This shows the result is true for $n = 1$ and in the case where all the irreducible factors of $f(x)$ have degree 1.

Assume now by induction that the theorem has been proved for any field F, isomorphism φ, and polynomial $f(x) \in F[x]$ of degree $< n$. Let $p(x)$ be an irreducible factor of $f(x)$ in $F[x]$ of degree at least 2 and let $p'(x)$ be the corresponding irreducible factor of $f'(x)$ in $F'[x]$. If $\alpha \in E$ is a root of $p(x)$ and $\beta \in E'$ is a root of $p'(x)$, then by Theorem 8 we can extend φ to an isomorphism $\sigma' : F(\alpha) \xrightarrow{\sim} F'(\beta)$:

$$\begin{array}{ccccc} \sigma': & F(\alpha) & \xrightarrow{\sim} & F'(\beta) \\ & | & & | \\ \varphi: & F & \xrightarrow{\sim} & F'. \end{array}$$

Let $F_1 = F(\alpha)$, $F_1' = F'(\beta)$, so that we have the isomorphism $\sigma' : F_1 \xrightarrow{\sim} F_1'$. We have $f(x) = (x - \alpha)f_1(x)$ over F_1 where $f_1(x)$ has degree $n - 1$ and $f'(x) = (x - \beta)f_1'(x)$. The field E is a splitting field for $f_1(x)$ over F_1: all the roots of $f_1(x)$ are in E and if they were contained in any smaller extension L containing F_1, then, since F_1 contains α, L would also contain all the roots of $f(x)$, which would contradict the minimality of E as the splitting field of $f(x)$ over F. Similarly E' is a splitting field for $f_1'(x)$ over F_1'. Since the degrees of $f_1(x)$ and $f_1'(x)$ are less than n, by induction there exists a map $\sigma : E \xrightarrow{\sim} E'$ extending the isomorphism $\sigma' : F_1 \xrightarrow{\sim} F_1'$. This gives the extended

diagram:

$$\begin{array}{lccc} \sigma: & E & \xrightarrow{\sim} & E' \\ & | & & | \\ \sigma': & F_1 & \xrightarrow{\sim} & F_1' \\ & | & & | \\ \varphi: & F & \xrightarrow{\sim} & F'. \end{array}$$

Then as the diagram indicates, σ restricted to F_1 is the isomorphism σ', so in particular σ restricted to F is σ' restricted to F, which is φ, showing that σ is an extension of φ, completing the proof.

Corollary 28. (Uniqueness of Splitting Fields) Any two splitting fields for a polynomial $f(x) \in F[x]$ over a field F are isomorphic.

Proof: Take φ to be the identity mapping from F to itself and E and E' to be two splitting fields for $f(x)(= f'(x))$.

As we mentioned before, this result justifies the terminology of *the* splitting field for $f(x)$ over F, since any two are isomorphic. Splitting fields play a natural role in the study of algebraic elements (if you are adjoining one root of a polynomial, why not adjoin *all* the roots?) and so take a particularly important role in Galois Theory.

We end this section with a discussion of field extensions of F which contain all the roots of *all* polynomials over F.

Definition. The field $\overline{F}$ is called an *algebraic closure* of F if $\overline{F}$ is algebraic over F and if every polynomial $f(x) \in F[x]$ splits completely over $\overline{F}$ (so that $\overline{F}$ can be said to contain all the elements algebraic over F).

Definition. A field K is said to be *algebraically closed* if every polynomial with coefficients in K has a root in K.

It is not obvious that algebraically closed fields exist nor that there exists an algebraic closure of a given field F (we shall prove this shortly).

Note that if K is algebraically closed, then in fact every $f(x) \in K[x]$ has *all* its roots in K, since by definition $f(x)$ has a root $\alpha \in K$, hence has a factor $x - \alpha$ in $K[x]$. The remaining factor of $f(x)$ then is a polynomial in $K[x]$, hence has a root, so has a linear factor etc., so that $f(x)$ must split completely. Hence if K is algebraically closed, then K itself is an algebraic closure of K and the converse is obvious, so that $K = \overline{K}$ if and only if K is algebraically closed.

The next result shows that the process of "taking the algebraic closure" actually stops after one step — taking the algebraic closure of an algebraic closure does not give a larger field: the field is already algebraically closed (notationally: $\overline{(\overline{F})} = \overline{F}$).

Proposition 29. Let $\overline{F}$ be an algebraic closure of F. Then $\overline{F}$ is algebraically closed.

Proof: Let $f(x)$ be a polynomial in $\overline{F}[x]$ and let α be a root of $f(x)$. Then α generates an algebraic extension $\overline{F}(\alpha)$ of $\overline{F}$, and $\overline{F}$ is algebraic over F. By Theorem 20, $\overline{F}(\alpha)$ is algebraic over F so in particular its element α is algebraic over F. But then $\alpha \in \overline{F}$, showing $\overline{F}$ is algebraically closed.

Given a field F we have already shown how to construct (finite) extensions of F containing all the roots of any given polynomial $f(x) \in F[x]$. Intuitively, an algebraic closure of F is given by the field "generated" by all of these fields. The difficulty with this is "generated" *where*?, since they are not all subfields of a given field. For a *finite* collection of polynomials $f_1(x), ..., f_k(x)$, we can identify their splitting fields as subfields of the splitting field of the product polynomial $f_1(x)...f_k(x)$, but the same idea used for an *infinite* number of polynomials requires numerous "bookkeeping" identifications and an application of Zorn's Lemma.

We shall instead construct an algebraic closure of F by first constructing an algebraically closed field containing F. The proof uses a clever idea of Artin which very neatly solves the "bookkeeping" problem of constructing a field containing the appropriate roots of polynomials (which also ultimately relies on Zorn's Lemma) by introducing a separate variable for every polynomial.

Proposition 30. For any field F there exists an algebraically closed field K containing F.

Proof: For every monic polynomial $f = f(x)$ with coefficients in F, let x_f denote an indeterminate and consider the polynomial ring $F[..., x_f, ...]$ generated over F by the variables x_f. In this polynomial ring consider the ideal I generated by the polynomials $f(x_f)$. If this ideal is not proper, then 1 is an element of the ideal, hence we have a relation

$$g_1 f_1(x_{f_1}) + g_2 f_2(x_{f_2}) + ... + g_n f_n(x_{f_n}) = 1$$

where the g_i, $i = 1, 2, ..., n$, are polynomials in the x_f. For $i = 1, 2, \ldots, n$ let $x_{f_i} = x_i$ and let $x_{n+1}, \ldots, x_m$ be the remaining variables occurring in the polynomials g_j, $j = 1, 2, \ldots, n$. Then the relation above reads

$$g_1(x_1, x_2, ..., x_m) f_1(x_1) + g_2(x_1, x_2, ..., x_m) f_2(x_2) + ... + g_n(x_1, x_2, ..., x_m) f_n(x_n) = 1.$$

Let F' be a finite extension of F containing a root α_i of $f_i(x)$ for $i = 1, 2, ..., n$. Letting $x_i = \alpha_i$, $i = 1, 2, \ldots, n$ and setting $x_{n+1} = \cdots = x_m = 0$, say, in the polynomial equation above would imply that $0 = 1$ in F', clearly impossible.

Since the ideal I is a proper ideal, it is contained in a maximal ideal $\mathcal{M}$ (this is where Zorn's Lemma is used). Then the quotient

$$K_1 = F[..., x_f, ...]/\mathcal{M}$$

is a field containing (an isomorphic copy of) F. Each of the polynomials f has a root in K_1 by construction, namely the image of x_f, since $f(x_f) \in I \subseteq \mathcal{M}$. We have constructed a field K_1 in which every polynomial with coefficients from F has a root. Performing the same construction with K_1 instead of F gives a field K_2 containing

K_1 in which all polynomials with coefficients from K_1 have a root. Continuing in this fashion we obtain a sequence of fields

$$F = K_0 \subseteq K_1 \subseteq K_2 \subseteq \ldots \subseteq K_j \subseteq K_{j+1} \subseteq \ldots$$

where every polynomial with coefficients in K_j has a root in K_{j+1}, $j = 0, 1, \ldots$. Let

$$K = \bigcup_{j \geq 0} K_j$$

be the union of these fields. Then K is clearly a field containing F. Since K is the union of the fields K_j, the coefficients of any polynomial $h(x)$ in $K[x]$ all lie in some field K_N for N sufficiently large. But then $h(x)$ has a root in K_{N+1}, so has a root in K. It follows that K is algebraically closed, completing the proof.

We now use the algebraically closed field containing F to construct an algebraic closure of F:

Proposition 31. Let K be an algebraically closed field and let F be a subfield of K. Then the collection of elements $\overline{F}$ of K which are algebraic over F is an algebraic closure of F. An algebraic closure of F is unique up to isomorphism.

Proof: By definition, $\overline{F}$ is an algebraic extension of F. Every polynomial $f(x) \in F[x]$ splits completely over K into linear factors $x - \alpha$ (the same is true for every polynomial even in $K[x]$). But each α is a root of $f(x)$, so is algebraic over F, hence is an element of $\overline{F}$. It follows that all the linear factors $x - \alpha$ have coefficients in $\overline{F}$, i.e. $f(x)$ splits completely in $\overline{F}[x]$ and $\overline{F}$ is an algebraic closure of F.

The uniqueness (up to isomorphism) of the algebraic closure is natural in light of the uniqueness (up to isomorphism) of splitting fields, and is proved along the same lines together with an application of Zorn's Lemma and will be omitted.

We shall prove later using Galois theory the following result (purely analytic proofs using complex analysis also exist).

Theorem. (Fundamental Theorem of Algebra) The field $\mathbb{C}$ is algebraically closed.

By Proposition 31, we immediately obtain:

Corollary 32. The field $\mathbb{C}$ contains an algebraic closure for any of its subfields. In particular, $\overline{\mathbb{Q}}$, the collection of complex numbers algebraic over $\mathbb{Q}$, is an algebraic closure of $\mathbb{Q}$.

The point of these considerations is that all the computations involving elements algebraic over a field F may be viewed as taking place in one (large) field, namely $\overline{F}$. Similarly, we can speak sensibly of the composite of any collection of algebraic extensions by viewing them all as subfields of an algebraic closure. In the case of $\mathbb{Q}$ or finite extensions of $\mathbb{Q}$ we may consider all of our computations as occurring in $\mathbb{C}$.

EXERCISES

1. Determine the splitting field over $\mathbb{Q}$ for $x^4 - 2$.
2. Determine the splitting field over $\mathbb{Q}$ of $x^4 + 2$.
3. Determine the splitting field over $\mathbb{Q}$ for $x^4 + x^2 + 1$.
4. Determine the splitting field over $\mathbb{Q}$ of $x^6 - 4$.

13.5 SEPARABLE AND INSEPARABLE EXTENSIONS

Let F be a field and let $f(x) \in F[x]$ be a polynomial. Over a splitting field for $f(x)$ we have the factorization

$$f(x) = (x - \alpha_1)^{n_1}(x - \alpha_2)^{n_2}\ldots(x - \alpha_k)^{n_k}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_k$ are distinct elements of the splitting field and $n_i \geq 1$ for $i = 1, 2, \ldots, k$. Recall that α_i is called a *multiple* root if $n_i > 1$ and is called a *simple* root if $n_i = 1$. The integer n_i is called the *multiplicity* of the root α_i.

Definition. A polynomial over F is called *separable* if it has no multiple roots (i.e. all its roots are distinct). A polynomial which is not separable is called *inseparable*.

Note that if a polynomial $f(x)$ has distinct roots in one splitting field then $f(x)$ has distinct roots in any splitting field (since this is equivalent to $f(x)$ factoring into distinct linear factors and the two splitting fields are isomorphic over F), so that we need not specify the field containing all the roots of $f(x)$.

Examples

(1) The polynomial $x^2 - 2$ is separable over $\mathbb{Q}$ since its two roots $\pm\sqrt{2}$ are distinct. The polynomial $(x^2 - 2)^n$ for any $n \geq 2$ is inseparable since it has the multiple roots $\pm\sqrt{2}$, each with multiplicity n.

(2) The polynomial $x^2 - t$ $(= x^2 + t)$ over the field $F = \mathbb{F}_2(t)$ of rational functions in t with coefficients from $\mathbb{F}_2$ is irreducible as we've seen before, but is not separable. If $\sqrt{t}$ denotes a root in some extension field (note that $\sqrt{t} \notin F$), then

$$(x - \sqrt{t})^2 = x^2 - 2x\sqrt{t} + t = x^2 + t = x^2 - t$$

since F is a field of characteristic 2. Hence this irreducible polynomial has only one root (with multiplicity 2), so is not separable over F.

There is a simple criterion to check whether a polynomial has multiple roots.

Definition. The *derivative* of the polynomial

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \in F[x]$$

is defined to be the polynomial

$$D_x f(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \ldots + 2a_2 x + a_1 \in F[x].$$

This formula is nothing but the usual formula for the derivative of a polynomial familiar from calculus. It is purely algebraic and so can be applied to a polynomial over an arbitrary field F, where the analytic notion of derivative (involving limits – a *continuous* operation) may not exist.

The usual (calculus) formulas for derivatives hold for derivatives in this situation as well, for example the formulas for the derivative of a sum and of a product:

$$D_x(f(x) + g(x)) = D_x f(x) + D_x g(x)$$
$$D_x(f(x)g(x)) = f(x)D_x g(x) + (D_x f(x))g(x).$$

These formulas can be proved directly from the definition for polynomials and do not require any limiting operations and are left as an exercise.

Proposition 33. A polynomial $f(x)$ has a multiple root α if and only if α is also a root of $D_x f(x)$, i.e. $f(x)$ and $D_x f(x)$ are both divisible by the minimal polynomial for α. In particular, $f(x)$ is separable if and only if it is relatively prime to its derivative: $(f(x), D_x f(x)) = 1$.

Proof: Suppose first that α is a multiple root of $f(x)$. Then over a splitting field,

$$f(x) = (x - \alpha)^n g(x)$$

for some integer $n \geq 2$ and some polynomial $g(x)$. Taking derivatives we obtain

$$D_x f(x) = n(x - \alpha)^{n-1} g(x) + (x - \alpha)^n D_x g(x)$$

which shows ($n \geq 2$) that $D_x f(x)$ has α as a root.

Conversely, suppose that α is a root of both $f(x)$ and $D_x f(x)$. Then write

$$f(x) = (x - \alpha)h(x)$$

for some polynomial $h(x)$ and take the derivative:

$$D_x f(x) = h(x) + (x - \alpha)D_x h(x).$$

Since $D_x f(\alpha) = 0$ by assumption, substituting α into the last equation shows that $h(\alpha) = 0$. Hence $h(x) = (x - \alpha)h_1(x)$ for some polynomial $h_1(x)$, and

$$f(x) = (x - \alpha)^2 h_1(x)$$

showing that α is a multiple root of $f(x)$.

The equivalence with divisibility by the minimal polynomial for α follows from Proposition 9. The last statement is then clear (let α denote any root of a common factor of $f(x)$ and $D_x f(x)$).

Examples

(1) The polynomial $x^{p^n} - x$ over $\mathbb{F}_p$ has derivative $p^n x^{p^n - 1} - 1 = -1$ since the field has characteristic p. Since in this case the derivative has no roots at all, it follows that the polynomial has no multiple roots, hence is separable.

(2) The polynomial $x^n - 1$ has derivative nx^{n-1}. Over any field of characteristic not dividing n (including characteristic 0) this polynomial has only the root 0 (of multiplicity $n - 1$), which is not a root of $x^n - 1$. Hence $x^n - 1$ is separable and there

are n distinct n^{th} roots of unity. We saw this directly over $\mathbb{Q}$ by exhibiting n distinct solutions over $\mathbb{C}$.

(3) If F is of characteristic p and p divides n, then there are fewer than n distinct n^{th} roots of unity over F: in this case the derivative is identically 0 since $n = 0$ in F. In fact *every* root of $x^n - 1$ is multiple in this case.

Corollary 34. Every irreducible polynomial over a field of characteristic 0 (for example, $\mathbb{Q}$) is separable. A polynomial over such a field is separable if and only if it is the product of distinct irreducible polynomials.

Proof: Suppose F is a field of characteristic 0 and $p(x) \in F[x]$ is irreducible of degree n. Then the derivative $D_x p(x)$ is a polynomial of degree $n - 1$. Up to constant factors the only factors of $p(x)$ in $F[x]$ are 1 and $p(x)$, so $D_x p(x)$ must be relatively prime to $p(x)$. This shows that any irreducible polynomial over a field of characteristic 0 is separable. The second statement of the corollary is then clear since distinct irreducibles never have zeros in common (by Proposition 9).

The point in the proof of the corollary that can fail in characteristic p is the statement that the derivative $D_x p(x)$ is of degree $n - 1$. In characteristic p the derivative of any power x^{pm} of x^p is identically 0:

$$D_x(x^{pm}) = pmx^{pm-1} = 0$$

so it is possible for the degree of the derivative to decrease by more than one. If the derivative $D_x p(x)$ of the *irreducible* polynomial $p(x)$ is nonzero, however, then just as before we conclude that $p(x)$ must be separable.

It is clear from the definition of the derivative that if $p(x)$ is a polynomial whose derivative is 0, then every power of x in $p(x)$ must be a multiple of p where p is the characteristic of F:

$$p(x) = a_m x^{mp} + a_{m-1}x^{(m-1)p} + \ldots + a_1 x^p + a_0.$$

Letting

$$p_1(x) = a_m x^m + a_{m-1}x^{m-1} + \ldots + a_1 x + a_0$$

we see that $p(x)$ is a polynomial in x^p, namely $p(x) = p_1(x^p)$.

We now prove a simple but important result about raising to the pth power in a field of characteristic p.

Proposition 35. Let F be a field of characteristic p. Then for any $a, b \in F$,

$$(a + b)^p = a^p + b^p$$
$$(ab)^p = a^p b^p.$$

Put another way, the map

$$\varphi : F \to F$$
$$a \mapsto a^p$$

is an injective field homomorphism.

Proof: The Binomial Theorem for expanding $(a+b)^n$ for any positive integer n holds (by the standard induction proof) over any commutative ring:

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \ldots + \binom{n}{i}a^{n-i}b^i + \ldots + b^n .$$

It should be observed that the binomial coefficients

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

are integers (recall that $m\alpha$ for $m \in \mathbb{Z}$ is defined for α an element of any ring) and here are elements of the prime field.

If p is a prime, then the binomial coefficients $\binom{p}{i}$ for $i = 1, 2, \ldots, p-1$ are all divisible by p since for these values of i the numbers $i!$ and $(p-i)!$ only involve factors smaller than p, hence are relatively prime to p and so cannot cancel the factor of p in the numerator of the expression $\dfrac{p!}{i!(p-i)!}$. It follows that over a field of characteristic p all the intermediate terms in the expansion of $(a+b)^p$ are 0, which gives the first equation of the proposition. The second equation is trivial, as is the fact that φ is injective.

Definition. The map in Proposition 35 is called the *Frobenius endomorphism* of F.

Corollary 36. Suppose that $\mathbb{F}$ is a finite field of characteristic p. Then every element of $\mathbb{F}$ is a pth power in $\mathbb{F}$ (notationally, $\mathbb{F} = \mathbb{F}^p$).

Proof: The injectivity of the Frobenius endomorphism of $\mathbb{F}$ implies that it is also surjective when $\mathbb{F}$ is finite, which is the statement of the corollary.

We now prove the analogue of Corollary 34 for finite fields.

Let $\mathbb{F}$ be a finite field and suppose that $p(x) \in \mathbb{F}[x]$ is an irreducible polynomial with coefficients in $\mathbb{F}$. If $p(x)$ were inseparable then we have seen that $p(x) = q(x^p)$ for some polynomial $q(x) \in \mathbb{F}[x]$. Let

$$q(x) = a_m x^m + a_{m-1}x^{m-1} + \ldots + a_1 x + a_0.$$

By Corollary 36, each a_i, $i = 1, 2, \ldots, m$ is a pth power in $\mathbb{F}$, say $a_i = b_i^p$. Then by Proposition 35 we have

$$\begin{aligned} p(x) = q(x^p) &= a_m(x^p)^m + a_{m-1}(x^p)^{m-1} + \ldots + a_1 x^p + a_0 \\ &= b_m^p(x^p)^m + b_{m-1}^p(x^p)^{m-1} + \ldots + b_1^p x^p + b_0^p \\ &= (b_m x^m)^p + (b_{m-1}x^{m-1})^p + \ldots + (b_1 x)^p + (b_0)^p \\ &= (b_m x^m + b_{m-1}x^{m-1} + \ldots + b_1 x + b_0)^p \end{aligned}$$

which shows that $p(x)$ is the pth power of a polynomial in $\mathbb{F}[x]$, a contradiction to the irreducibility of $p(x)$. This proves:

Proposition 37. Every irreducible polynomial over a finite field $\mathbb{F}$ is separable. A polynomial in $\mathbb{F}[x]$ is separable if and only if it is the product of distinct irreducible polynomials in $\mathbb{F}[x]$.

The important part of the proof of this result is the fact that every element in the characteristic p field $\mathbb{F}$ was a pth power in $\mathbb{F}$. This suggests the following definition:

Definition. A field K of characteristic p is called *perfect* if every element of K is a pth power in K: $K = K^p$. Any field of characteristic 0 is also called perfect.

With this definition, we see that we have proved that every irreducible polynomial over a perfect field is separable. It is not hard to see that if K is not perfect then there are inseparable irreducible polynomials.

Example: (Existence and Uniqueness of Finite Fields)

Let $n > 0$ be any positive integer and consider the splitting field of the polynomial $x^{p^n} - x$ over $\mathbb{F}_p$. We have already seen that this polynomial is separable, hence has precisely p^n roots. Let α and β be any two roots of this polynomial, so that $\alpha^{p^n} = \alpha$ and $\beta^{p^n} = \beta$. Then $(\alpha\beta)^{p^n} = \alpha\beta$, $(\alpha^{-1})^{p^n} = \alpha^{-1}$ and by Proposition 35 also

$$(\alpha + \beta)^{p^n} = \alpha^{p^n} + \beta^{p^n} = \alpha + \beta.$$

Hence the set $\mathbb{F}$ consisting of the p^n distinct roots of $x^{p^n} - x$ over $\mathbb{F}_p$ is *closed* under addition, multiplication and inverses in its splitting field. It follows that $\mathbb{F}$ is a subfield, hence in fact must be the splitting field. Since the number of elements is p^n, we have $[\mathbb{F} : \mathbb{F}_p] = n$, which shows that there exist finite fields of degree n over $\mathbb{F}_p$ for any $n > 0$.

Let now $\mathbb{F}$ be any finite field of characteristic p. If $\mathbb{F}$ is of dimension n over its prime subfield $\mathbb{F}_p$, then $\mathbb{F}$ has precisely p^n elements. Since the multiplicative group $\mathbb{F}^\times$ is (in fact cyclic) of order $p^n - 1$, we have $\alpha^{p^n-1} = 1$ for every $\alpha \neq 0$ in $\mathbb{F}$, so that $\alpha^{p^n} = \alpha$ for every $\alpha \in \mathbb{F}$. But this means α is a root of $x^{p^n} - x$, hence $\mathbb{F}$ is contained in a splitting field for this polynomial. Since we have seen that the splitting field has order p^n this shows that $\mathbb{F}$ is a splitting field for $x^{p^n} - x$. Since splitting fields are unique up to isomorphism, this proves that *finite fields of any order p^n exist and are unique up to isomorphism*. We shall denote the finite field of order p^n by $\mathbb{F}_{p^n}$.

We shall consider finite fields more later.

We now investigate further the structure of inseparable irreducible polynomials over fields of characteristic p. We have seen above that if $p(x)$ is an irreducible polynomial which is not separable, then its derivative $D_x p(x)$ is identically 0, so that $p(x) = p_1(x^p)$ for some polynomial $p_1(x)$. The polynomial $p_1(x)$ may or may not itself be separable. If not, then it too is a polynomial in x^p, $p_1(x) = p_2(x^p)$, so that $p(x)$ is a polynomial in x^{p^2}: $p(x) = p_2(x^{p^2})$. Continuing in this fashion we see that there is a uniquely defined power p^k of p such that $p(x) = p_k(x^{p^k})$ where $p_k(x)$ has nonzero derivative. It is clear that $p_k(x)$ is irreducible since any factorization of $p_k(x)$ would, after replacing x by x^{p^k}, immediately imply a factorization of the irreducible $p(x)$. It follows that $p_k(x)$ is separable. We summarize this as:

Proposition 38. Let $p(x)$ be an irreducible polynomial over a field F of characteristic p. Then there is a unique integer $k \geq 0$ and a unique irreducible separable polynomial $p_{sep}(x) \in F[x]$ such that

$$p(x) = p_{sep}(x^{p^k}).$$

Definition. Let $p(x)$ be an irreducible polynomial over a field of characteristic p. The degree of $p_{sep}(x)$ in the last proposition is called the *separable degree* of $p(x)$, denoted $\deg_s p(x)$. The integer p^k in the proposition is called the *inseparable degree* of $p(x)$, denoted $\deg_i p(x)$.

From the definitions and the proposition we see that $p(x)$ is separable if and only if its inseparability degree is 1 if and only if its degree is equal to its separable degree. Also, computing degrees in the relation $p(x) = p_{sep}(x^{p^k})$ we see that

$$\deg p(x) = \deg_s p(x) \deg_i p(x).$$

Examples

(1) The polynomial $p(x) = x^2 - t$ over $F = \mathbb{F}_2(t)$ considered above has derivative 0, hence is not separable (as we determined earlier). Here $p_{sep}(x) = x - t$ with inseparability degree 2.

(2) The polynomial $p(x) = x^{2^m} - t$ over $F = \mathbb{F}_2(t)$ is irreducible with the same separable polynomial part, but with inseparability degree 2^m.

(3) The polynomial $(x^{p^2} - t)(x^p - t)$ over $F = \mathbb{F}_p(t)$ has (two) inseparable irreducible factors so is inseparable. This polynomial cannot be written in the form $f_{sep}(x^{p^k})$ where $f_{sep}(x)$ is separable, which is the reason we restricted to *irreducible* polynomials above. This example also shows that there is no analogous factorization to define the separable and inseparable degrees of a general polynomial.

The notion of separability carries over to the fields generated by the roots of these polynomials.

Definition. The field K is said to be *separable* (or *separably algebraic*) over F if every element of K is the root of a separable polynomial over F (equivalently, the minimal polynomial over F of every element of K is separable). A field which is not separable is *inseparable*.

We have seen that the issue of separability is straightforward for finite extensions of perfect fields since for these fields the minimal polynomial of an algebraic element is irreducible hence separable.

Corollary 39. Every finite extension of a perfect field is separable. In particular, every finite extension of either $\mathbb{Q}$ or a finite field is separable.

We shall consider separable and inseparable extensions more after developing some Galois Theory, in particular defining the separable and inseparable *degree* of the extension K/F.

EXERCISES

1. Prove that the derivative D_x of a polynomial satisfies $D_x(f(x)+g(x)) = D_x(f(x)) + D_x(g(x))$ and $D_x(f(x)g(x)) = D_x(f(x))g(x) + D_x(g(x))f(x)$ for any two polynomials $f(x)$ and $g(x)$.
2. Find all irreducible polynomials of degrees 1,2 and 4 over $\mathbb{F}_2$ and prove that their product is $x^{16} - x$.
3. Prove that d divides n if and only if $x^d - 1$ divides $x^n - 1$. [Note that if $n = qd + r$ then $x^n - 1 = (x^{qd+r} - x^r) + (x^r - 1)$.]
4. Let $a > 1$ be an integer. Prove for any positive integers n, d that d divides n if and only if $a^d - 1$ divides $a^n - 1$ (cf. the previous exercise). Conclude in particular that $\mathbb{F}_{p^d} \subseteq \mathbb{F}_{p^n}$ if and only if d divides n.
5. For any prime p and any nonzero $a \in \mathbb{F}_p$ prove that $x^p - x + a$ is irreducible and separable over $\mathbb{F}_p$. [For the irreducibility: One approach — prove first that if α is a root then $\alpha + 1$ is also a root. Another approach — suppose it's reducible and compute derivatives.]
6. Prove that $x^{p^n-1} - 1 = \prod_{\alpha \in \mathbb{F}_{p^n}^\times} (x - \alpha)$. Conclude that $\prod_{\alpha \in \mathbb{F}_{p^n}^\times} \alpha = (-1)^{p^n}$ so the product of the nonzero elements of a finite field is $+1$ if $p = 2$ and -1 if p is odd. For p odd and $n = 1$ derive *Wilson's Theorem*: $(p-1)! \equiv -1 \pmod{p}$.
7. Suppose K is a field of characteristic p which is not a perfect field: $K \neq K^p$. Prove there exist irreducible inseparable polynomials over K. Conclude that there exist inseparable finite extensions of K.
8. Prove that $f(x)^p = f(x^p)$ for any polynomial $f(x) \in \mathbb{F}_p[x]$.
9. Show that the binomial coefficient $\binom{pn}{pi}$ is the coefficient of x^{pi} in the expansion of $(1+x)^{pn}$. Working over $\mathbb{F}_p$ show that this is the coefficient of $(x^p)^i$ in $(1+x^p)^n$ and hence prove that $\binom{pn}{pi} \equiv \binom{n}{i} \pmod{p}$.
10. Let $f(x_1, x_2, \ldots, x_n) \in \mathbb{Z}[x_1, x_2, \ldots, x_n]$ be a polynomial in the variables $x_1, x_2, \ldots, x_n$ with integer coefficients. For any prime p prove that the polynomial

$$f(x_1, x_2, \ldots, x_n)^p - f(x_1^p, x_2^p, \ldots, x_n^p) \in \mathbb{Z}[x_1, x_2, \ldots, x_n]$$

has all its coefficients divisible by p.

13.6 CYCLOTOMIC POLYNOMIALS AND EXTENSIONS

The purpose of this section is to prove that the cyclotomic extension

$$\mathbb{Q}(\zeta_n)/\mathbb{Q}$$

generated by the nth roots of unity over $\mathbb{Q}$ introduced in Section 4 is of degree $\varphi(n)$ where φ denotes Euler's phi-function (= the number of integers a, $1 \leq a < n$ relatively prime to n = the order of the group $(\mathbb{Z}/n\mathbb{Z})^\times$).

Definition. Let μ_n denote the *group of* n^{th} *roots of unity over* $\mathbb{Q}$.

Then as we have already observed, $\mathbb{Z}/n\mathbb{Z} \cong \mu_n$ as groups (under multiplication on the right, addition on the left), given explicitly by the map $a \mapsto (\zeta_n)^a$ for a fixed

primitive n^{th} root of unity. The primitive n^{th} roots of unity are given by the residue classes prime to n so there are precisely $\varphi(n)$ primitive n^{th} roots of unity.

If d is a divisor of n and ζ is a dth root of unity, then ζ is also an nth root of unity since $\zeta^n = (\zeta^d)^{n/d} = 1$. Hence

$$\mu_d \subseteq \mu_n \qquad \text{for all } d \mid n.$$

Conversely, the order of any element of the group μ_n is a divisor of n so that if ζ is an nth root of unity which is also a dth root of unity for some smaller d then $d \mid n$.

Definition. Define the nth *cyclotomic polynomial* $\Phi_n(x)$ to be the polynomial whose roots are the primitive n^{th} roots of unity:

$$\Phi_n(x) = \prod_{\zeta \text{ primitive } \in \mu_n} (x - \zeta) = \prod_{\substack{1 \le a < n \\ (a,n)=1}} (x - \zeta_n{}^a)$$

(which is of degree $\varphi(n)$).

The roots of the polynomial $x^n - 1$ are precisely the nth roots of unity so we have the factorization

$$x^n - 1 = \prod_{\substack{\zeta^n = 1 \\ i.e.\ \zeta \in \mu_n}} (x - \zeta).$$

If we group together the factors $(x - \zeta)$ where ζ is an element of order d in μ_n (i.e. ζ is a primitive d^{th} root of unity) we obtain

$$x^n - 1 = \prod_{d \mid n} \prod_{\substack{\zeta \in \mu_d \\ \zeta \text{ primitive}}} (x - \zeta).$$

The inner product is $\Phi_d(x)$ by definition so we have the factorization

$$x^n - 1 = \prod_{d \mid n} \Phi_d(x). \tag{13.4}$$

Note incidentally that comparing degrees gives the identity

$$n = \sum_{d \mid n} \varphi(d).$$

This factorization allows us to compute $\Phi_n(x)$ for any n recursively: clearly $\Phi_1(x) = x - 1$ and $\Phi_2(x) = x + 1$. Then

$$x^3 - 1 = \Phi_1(x)\Phi_3(x) = (x - 1)\Phi_3(x)$$

which gives

$$\Phi_3(x) = x^2 + x + 1.$$

Similarly

$$x^4 - 1 = \Phi_1(x)\Phi_2(x)\Phi_4(x) = (x - 1)(x + 1)\Phi_4(x)$$

gives

$$\Phi_4(x) = x^2 + 1$$

(in these cases these could also be obtained directly from the explicit roots of unity). Continuing in this fashion we can compute $\Phi_n(x)$ for any n. Note also that for p a prime we recover our polynomial

$$\Phi_p(x) = x^{p-1} + x^{p-2} + \ldots + x + 1.$$

For some small values of n the polynomials are

$$\begin{aligned}
\Phi_5(x) &= x^4 + x^3 + x^2 + x + 1 \\
\Phi_6(x) &= x^2 - x + 1 \\
\Phi_7(x) &= x^6 + x^5 + x^4 + x^3 + x^2 + x + 1 \\
\Phi_8(x) &= x^4 + 1 \\
\Phi_9(x) &= x^6 + x^3 + 1 \\
\Phi_{10}(x) &= x^4 - x^3 + x^2 - x + 1 \\
\Phi_{11}(x) &= x^{10} + x^9 + \ldots + x + 1 \\
\Phi_{12}(x) &= x^4 - x^2 + 1.
\end{aligned}$$

For all the values computed above, $\Phi_n(x)$ was a (monic) polynomial with integer coefficients. This is always the case:

Lemma 40. The cyclotomic polynomial $\Phi_n(x)$ is a monic polynomial in $\mathbb{Z}[x]$ of degree $\varphi(n)$.

Proof: It is clear that $\Phi_n(x)$ is monic and has degree $\varphi(n)$. We must show the coefficients lie in $\mathbb{Z}$. We use induction on n. The result is true for $n = 1$ (and $n \leq 12$). Assume by induction that $\Phi_d(x) \in \mathbb{Z}[x]$ for all $1 \leq d < n$. Then $x^n - 1 = f(x)\Phi_n(x)$ where $f(x) = \prod_{\substack{d|n \\ d<n}} \Phi_d(x)$ is monic and has coefficients in $\mathbb{Z}$. Since $f(x)$ clearly divides $x^n - 1$ in $F[x]$ where $F = \mathbb{Q}(\zeta_n)$ is the field of nth roots of unity and both $f(x)$ and $x^n - 1$ have coefficients in $\mathbb{Q}$, $f(x)$ divides $x^n - 1$ in $\mathbb{Q}[x]$ by the Division Algorithm (cf. the remark at the end of Section 9.2). By Gauss' Lemma, $f(x)$ divides $x^n - 1$ in $\mathbb{Z}[x]$, hence $\Phi_n(x) \in \mathbb{Z}[x]$.

We remark in passing that while all the coefficients of $\Phi_n(x)$ in the examples computed above were $0, \pm 1$, it is known that there are cyclotomic polynomials with arbitrarily large coefficients.

Theorem 41. The cyclotomic polynomial $\Phi_n(x)$ is an irreducible monic polynomial in $\mathbb{Z}[x]$ of degree $\varphi(n)$.

Proof: We must show that $\Phi_n(x)$ is irreducible. If not then we have a factorization

$$\Phi_n(x) = f(x)g(x) \qquad \text{with } f(x), g(x) \text{ monic in } \mathbb{Z}[x]$$

where we take $f(x)$ to be an *irreducible* factor of $\Phi_n(x)$. Let ζ be a primitive n^{th} root of 1 which is a root of $f(x)$ (so then $f(x)$ is the minimal polynomial for ζ over $\mathbb{Q}$) and let p denote *any* prime not dividing n. Then ζ^p is again a primitive n^{th} root of 1, hence is a root of either $f(x)$ or $g(x)$.

Suppose $g(\zeta^p) = 0$. Then ζ is a root of $g(x^p)$ and since $f(x)$ is the minimal polynomial for ζ, $f(x)$ must divide $g(x^p)$ in $\mathbb{Z}[x]$, say

$$g(x^p) = f(x)h(x)\,, \qquad h(x) \in \mathbb{Z}[x].$$

If we reduce this equation mod p, we obtain

$$\bar{g}(x^p) = \bar{f}(x)\bar{h}(x) \qquad \text{in } \mathbb{F}_p[x].$$

By the remarks of the last section,

$$\bar{g}(x^p) = (\bar{g}(x))^p$$

so we have the equation

$$(\bar{g}(x))^p = \bar{f}(x)\bar{h}(x)$$

in the U.F.D. $\mathbb{F}_p[x]$. It follows that $\bar{f}(x)$ and $\bar{g}(x)$ have a factor in common in $\mathbb{F}_p[x]$.

Now, from $\Phi_n(x) = f(x)g(x)$ we see by reducing mod p that $\overline{\Phi}_n(x) = \bar{f}(x)\bar{g}(x)$, and so by the above it follows that $\overline{\Phi}_n(x) \in \mathbb{F}_p[x]$ has a multiple root. But then also $x^n - 1$ would have a multiple root over $\mathbb{F}_p$ since it has $\overline{\Phi}_n(x)$ as a factor. This is a contradiction since we have seen in the last section that there are n distinct roots of $x^n - 1$ over any field of characteristic not dividing n.

Hence ζ^p must be a root of $f(x)$. Since this applies to every root ζ of $f(x)$, it follows that ζ^a is a root of $f(x)$ for every integer a relatively prime to n: write $a = p_1p_2...p_k$ as a product of (not necessarily distinct) primes not dividing n so that ζ^{p_1} is a root of $f(x)$, so also $(\zeta^{p_1})^{p_2}$ is a root of $f(x)$, etc. But this means that *every* primitive n^{th} root of unity is a root of $f(x)$, i.e. $f(x) = \Phi_n(x)$, showing $\Phi_n(x)$ is irreducible.

Corollary 42. The degree over $\mathbb{Q}$ of the cyclotomic field of nth roots of unity is $\varphi(n)$:

$$[\mathbb{Q}(\zeta_n) : \mathbb{Q}] = \varphi(n).$$

Proof: By the theorem, $\Phi_n(x)$ is the minimal polynomial for any primitive n^{th} root of unity ζ_n.

Example

The cyclotomic field $\mathbb{Q}(\zeta_8)$ of the 8th roots of unity is of degree $\varphi(8) = 4$ over $\mathbb{Q}$. This field contains the 4th roots of unity, i.e. $\mathbb{Q}(i) \subset \mathbb{Q}(\zeta_8)$ as well as the element $\zeta_8 + \zeta_8{}^7 = \sqrt{2}$ (recall the explicit roots of unity in Section 4). It follows that

$$\mathbb{Q}(\zeta_8) = \mathbb{Q}(i, \sqrt{2}).$$

One interesting number-theoretic application of the cyclotomic polynomials outlined in the exercises is the proof that for any n there are infinitely many primes which are congruent to 1 modulo n. The complete factorization in $\mathbb{F}_p[x]$ of $\Phi_\ell(x)$ for a prime ℓ (which is irreducible in $\mathbb{Z}[x]$) is described in Exercise 8 below.

We shall return to the example of cyclotomic fields after we have developed some Galois Theory.

EXERCISES

1. Suppose m and n are relatively prime positive integers. Let ζ_m be a primitive m^{th} root of unity and let ζ_n be a primitive n^{th} root of unity. Prove that $\zeta_m\zeta_n$ is a primitive mn^{th} root of unity.
2. Let ζ_n be a primitive n^{th} root of unity and let d be a divisor of n. Prove that ζ_n^d is a primitive $(n/d)^{\text{th}}$ root of unity.
3. Prove that if a field contains the nth roots of unity for n odd then it also contains the $2n^{\text{th}}$ roots of unity.
4. Prove that if $n = p^k m$ where p is a prime and m is relatively prime to p then there are precisely m distinct n^{th} roots of unity over a field of characteristic p.
5. Let K be a finite extension of $\mathbb{Q}$. Prove there are only a finite number of roots of unity in K.
6. Prove that for n odd, $n > 1$, $\Phi_{2n}(x) = \Phi_n(-x)$.
7. Use the Möbius Inversion formula indicated in Section 14.3 to prove
$$\Phi_m(x) = \prod_{d|n}(x^d - 1)^{\mu(m/d)}.$$
8. Let ℓ be a prime and let $\Phi_\ell(x) = \frac{x^\ell - 1}{x-1} = x^{\ell-1} + x^{\ell-2} + \ldots + x + 1 \in \mathbb{Z}[x]$ be the ℓth cyclotomic polynomial, which is irreducible over $\mathbb{Z}$ by Proposition 41. This exercise determines the factorization of $\Phi_\ell(x)$ modulo p for any prime p. Let ζ denote any fixed primitive ℓ^{th} root of unity.
   (a) Show that if $p = l$ then $\Phi_\ell(x) = (x-1)^{\ell-1} \in \mathbb{F}_\ell[x]$.
   (b) Suppose $p \neq \ell$ and let f denote the order of p mod ℓ, i.e. f is the smallest power of p with $p^f \equiv 1$ mod ℓ. Use the fact that $\mathbb{F}_{p^n}^\times$ is a cyclic group to show that $n = f$ is the smallest power p^n of p with $\zeta \in \mathbb{F}_{p^n}$. Conclude that the minimal polynomial of ζ over $\mathbb{F}_p$ has degree f.
   (c) Show that $\mathbb{F}_p(\zeta) = \mathbb{F}_p(\zeta^a)$ for any integer a not divisible by ℓ. [One inclusion is obvious. For the other, note that $\zeta = (\zeta^a)^b$ where b is the multiplicative inverse of a mod ℓ.] Conclude using (b) that, in $\mathbb{F}_p[x]$, $\Phi_\ell(x)$ is the product of $\frac{\ell-1}{f}$ distinct irreducible polynomials of degree f.
   (d) In particular, prove that, viewed in $\mathbb{F}_p[x]$, $\Phi_7(x) = x^6 + x^5 + \ldots + x + 1$ is $(x-1)^6$ for $p = 7$, a product of distinct linear factors for $p \equiv 1$ mod 7, a product of 3 irreducible quadratics for $p \equiv 6$ mod 7, a product of 2 irreducible cubics for $p \equiv 2, 4$ mod 7, and is irreducible for $p \equiv 3, 5$ mod 7.
9. Suppose A is an $n \times n$ matrix over $\mathbb{C}$ for which $A^k = I$ for some integer $k \geq 1$. Show that A can be diagonalized. Show that the matrix $A = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}$ where α is an element of a field of characteristic p satisfies $A^p = I$ and cannot be diagonalized if $\alpha \neq 0$.
10. Let φ denote the Frobenius map $x \mapsto x^p$ on the finite field $\mathbb{F}_{p^n}$. Prove that φ gives an isomorphism of $\mathbb{F}_{p^n}$ to itself (such an isomorphism is called an *automorphism*). Prove that φ^n is the identity map and that no lower power of φ is the identity.

11. Let φ denote the Frobenius map $x \mapsto x^p$ on the finite field $\mathbb{F}_{p^n}$ as in the previous exercise. Determine the rational canonical form over $\mathbb{F}_p$ for φ considered as an $\mathbb{F}_p$-linear transformation of the n-dimensional $\mathbb{F}_p$-vector space $\mathbb{F}_{p^n}$.

12. Let φ denote the Frobenius map $x \mapsto x^p$ on the finite field $\mathbb{F}_{p^n}$ as in the previous exercise. Determine the Jordan canonical form (over a field containing all the eigenvalues) for φ considered as an $\mathbb{F}_p$-linear transformation of the n-dimensional $\mathbb{F}_p$-vector space $\mathbb{F}_{p^n}$.

13. (*Wedderburn's Theorem on finite division rings*) This exercise outlines a proof (following Witt) of Wedderburn's Theorem that a finite division ring D is a field (i.e. is commutative).
    (a) Let Z denote the center of D (i.e. the elements of D which commute with every element of D). Prove that Z is a field containing $\mathbb{F}_p$ for some prime p. If $Z = \mathbb{F}_q$ prove that D has order q^n for some integer n [D is a vector space over Z].
    (b) The nonzero elements $D^\times$ of D form a multiplicative group. For any $x \in D^\times$ show that the elements of D which commute with x form a division ring which contains Z. Show that this division ring is of order q^m for some integer m and that $m < n$ if x is not an element of Z.
    (c) Show that the class equation (Theorem 4.7) for the group $D^\times$ is

$$q^n - 1 = (q-1) + \sum_{i=1}^{r} \frac{q^n - 1}{|C_{D^\times}(x_i)|}$$

    where $x_1, x_2, \ldots, x_r$ are representatives of the distinct conjugacy classes in $D^\times$ not contained in the center of $D^\times$. Conclude from (b) that for each i, $|C_{D^\times}(x_i)| = q^{m_i} - 1$ for some $m_i < n$.
    (d) Prove that since $\dfrac{q^n - 1}{q^{m_i} - 1}$ is an integer (namely, the index $|D^\times : C_{D^\times}(x_i)|$) then m_i divides n (cf. Exercise 4 of Section 5). Conclude that $\Phi_n(x)$ divides $(x^n - 1)/(x^{m_i} - 1)$ and hence that the integer $\Phi_n(q)$ divides $(q^n - 1)/(q^{m_i} - 1)$ for $i = 1, 2, \ldots, r$.
    (e) Prove that (c) and (d) imply that $\Phi_n(q) = \prod_{\zeta \text{ primitive}}(q - \zeta)$ divides $q - 1$. Prove that $|q - \zeta| > q - 1$ (complex absolute value) for any root of unity $\zeta \neq 1$ [note that 1 is the closest point on the unit circle in $\mathbb{C}$ to the point q on the real line]. Conclude that $n = 1$, i.e. that $D = Z$ is a field.

The following exercises provide a proof that for any positive integer m there are infinitely many primes p with $p \equiv 1 \pmod{m}$. This is a special case of *Dirichlet's Theorem on Primes in Arithmetic Progressions* which states more generally that there are infinitely many primes p with $p \equiv a \pmod{m}$ for any a relatively prime to m.

14. Given any monic polynomial $P(x) \in \mathbb{Z}[x]$ of degree at least one show that there are infinitely many distinct prime divisors of the integers

$$P(1), P(2), P(3), \ldots, P(n), \ldots .$$

[Suppose the only primes dividing the values $P(n)$, $n = 1, 2, \ldots$, are the primes $p_1, p_2, \ldots, p_k$. Let N be an integer with $P(N) = a \neq 0$. Show that $Q(x) = a^{-1}P(N + a\,p_1p_2\ldots p_k\,x)$ is an element of $\mathbb{Z}[x]$ and that $Q(n) \equiv 1 \pmod{p_1p_2\ldots p_k}$ for $n = 1, 2, \ldots$. Conclude that there is some integer M such that $Q(M)$ has a prime factor different from $p_1, p_2, \ldots, p_k$ and hence that $P(N + ap_1p_2\ldots p_kM)$ has a prime factor different from $p_1, p_2, \ldots, p_k$.]

15. Let p be an odd prime not dividing m and let $\Phi_m(x)$ be the mth cyclotomic polynomial. Suppose $a \in \mathbb{Z}$ satisfies $\Phi_m(a) \equiv 0 \pmod{p}$. Prove that a is relatively prime to p and that the order of a in $(\mathbb{Z}/p\mathbb{Z})^\times$ is precisely m. [Since

$$x^m - 1 = \prod_{d|m} \Phi_d(x) = \Phi_m(x) \prod_{\substack{d|m \\ d<m}} \Phi_d(x)$$

we see first that $a^m - 1 \equiv 0 \pmod{p}$ i.e. $a^m \equiv 1 \pmod{p}$. If the order of a mod p were less than m, then $a^d \equiv 1 \pmod{p}$ for some d dividing m, so then $\Phi_d(a) \equiv 0 \pmod{p}$ for some $d < m$. But then $x^m - 1$ would have a as a multiple root mod p, a contradiction.]

16. Let $a \in \mathbb{Z}$. Show that if p is an odd prime dividing $\Phi_m(a)$ then either p divides m or $p \equiv 1 \pmod{m}$.

17. Prove there are infinitely many primes p with $p \equiv 1 \pmod{m}$.

14

Galois Theory

14.1 BASIC DEFINITIONS

In the previous chapter we proved the existence of a finite extension of a field F which contains all the roots of a given polynomial $f(x)$ whose coefficients are in F. The main idea of Galois Theory (named for Evariste Galois, 1811–1832) is to consider the relation of the group of permutations of the roots of $f(x)$ to the algebraic structure of its splitting field. The connection is given by the Fundamental Theorem of the next section. It can be viewed as another (extremely elegant) application of the important idea in mathematics that one (in our case algebraic) object *acting* on another provides structural information about both.

In this section we introduce the terminology and basic properties of the objects of interest. Let K be a field.

Definition

(1) An isomorphism σ of K with itself is called an *automorphism* of K. The collection of automorphisms of K is denoted $\mathrm{Aut}(K)$. If $\alpha \in K$ we shall write $\sigma\alpha$ for $\sigma(\alpha)$.

(2) An automorphism $\sigma \in \mathrm{Aut}(K)$ is said to *fix* an element $\alpha \in K$ if $\sigma\alpha = \alpha$. If F is a subset of K (for example, a subfield), then an automorphism σ is said to *fix* F if it fixes all the elements of F, i.e. $\sigma a = a$ for all $a \in F$.

Note that any field has at least one automorphism, the identity map, denoted by 1 and sometimes called the *trivial* automorphism.

The prime field of K is generated by $1 \in K$ and since any automorphism σ takes 1 to 1 (and 0 to 0), i.e. $\sigma(1) = 1$, it follows that $\sigma a = a$ for all a in the prime field. Hence any automorphism of a field K fixes its prime subfield. In particular we see that $\mathbb{Q}$ and $\mathbb{F}_p$ have only the trivial automorphism: $\mathrm{Aut}(\mathbb{Q}) = \{1\}$ and $\mathrm{Aut}(\mathbb{F}_p) = \{1\}$.

Definition. Let K/F be an extension of fields. Let $\mathrm{Aut}(K/F)$ be the collection of automorphisms of K which fix F.

Note that if F is the prime subfield of K then $\mathrm{Aut}(K) = \mathrm{Aut}(K/F)$ since every automorphism of K automatically fixes F.

If σ and τ are automorphisms of K then the composite $\sigma\tau$ (and also the composite $\tau\sigma$, which may not be the same) is defined and is again an automorphism of K.

Proposition 1. $\mathrm{Aut}(K)$ is a group under composition and $\mathrm{Aut}(K/F)$ is a subgroup.

Proof: It is clear that $\mathrm{Aut}(K)$ is a group. If σ and τ are automorphisms of K which fix F then also $\sigma\tau$ and σ^{-1} are the identity on F, which shows that $\mathrm{Aut}(K/F)$ is a subgroup.

The following proposition is extremely useful for determining the automorphisms of algebraic extensions.

Proposition 2. Let K/F be a field extension and let $\alpha \in K$ be algebraic over F. Then for any $\sigma \in \mathrm{Aut}(K/F)$, $\sigma\alpha$ is a root of the minimal polynomial for α over F i.e. $\mathrm{Aut}(K/F)$ permutes the roots of irreducible polynomials. Equivalently, any polynomial with coefficients in F having α as a root also has $\sigma\alpha$ as a root.

Proof: Suppose α satisfies the equation

$$\alpha^n + a_{n-1}\alpha^{n-1} + \ldots + a_1\alpha + a_0 = 0$$

where $a_0, a_1, \ldots, a_{n-1}$ are elements of F. Applying the automorphism σ we obtain (using the fact that σ is an additive homomorphism)

$$\sigma(\alpha^n) + \sigma(a_{n-1}\alpha^{n-1}) + \ldots + \sigma(a_1\alpha) + \sigma(a_0) = \sigma(0) = 0.$$

Using the fact that σ is also a multiplicative homomorphism this becomes

$$(\sigma(\alpha))^n + \sigma(a_{n-1})(\sigma(\alpha))^{n-1} + \ldots + \sigma(a_1)(\sigma(\alpha)) + \sigma(a_0) = 0.$$

By assumption, σ fixes all the elements of F, so $\sigma(a_i) = a_i$, $i = 0, 1, \ldots, n-1$. Hence

$$(\sigma\alpha)^n + a_{n-1}(\sigma\alpha)^{n-1} + \ldots + a_1(\sigma\alpha) + a_0 = 0.$$

But this says precisely that $\sigma\alpha$ is a root of the same polynomial over F as α. This proves the proposition.

Examples

(1) Let $K = \mathbb{Q}(\sqrt{2})$. If $\tau \in \mathrm{Aut}(\mathbb{Q}(\sqrt{2})) = \mathrm{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q})$, then $\tau(\sqrt{2}) = \pm\sqrt{2}$ since these are the two roots of the minimal polynomial for $\sqrt{2}$. Since τ fixes $\mathbb{Q}$, this determines τ completely:

$$\tau(a + b\sqrt{2}) = a \pm b\sqrt{2}.$$

The map $\sqrt{2} \mapsto \sqrt{2}$ is just the identity automorphism 1 of $\mathbb{Q}(\sqrt{2})$. The map $\sigma : \sqrt{2} \mapsto -\sqrt{2}$ is the isomorphism considered in Example 2 following Corollary 13.7. Hence $\mathrm{Aut}(\mathbb{Q}(\sqrt{2})) = \mathrm{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) = \{1, \sigma\}$ is a cyclic group of order 2 generated by σ.

(2) Let $K = \mathbb{Q}(\sqrt[3]{2})$. As before, if $\tau \in \mathrm{Aut}(K/\mathbb{Q})$, then τ is completely determined by its action on $\sqrt[3]{2}$ since

$$\tau(a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2) = a + b\tau\sqrt[3]{2} + c(\tau\sqrt[3]{2})^2.$$

Since $\tau\sqrt[3]{2}$ must be a root of $x^3 - 2$ and the other two roots of this equation are not elements of K (recall the splitting field of this polynomial is degree 6 over $\mathbb{Q}$), the only possibility is $\tau\sqrt[3]{2} = \sqrt[3]{2}$ i.e. $\tau = 1$. Hence $\mathrm{Aut}(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}) = 1$ is the trivial group.

In general, if K is generated over F by some collection of elements, then any automorphism $\sigma \in \mathrm{Aut}(K/F)$ is completely determined by what it does to the generators. If K/F is finite then K is finitely generated over F by algebraic elements so by the proposition the number of automorphisms of K fixing F is finite, i.e. $\mathrm{Aut}(K/F)$ is a finite group. In particular, the automorphisms of a finite extension can be considered as permutations of the roots of a finite number of equations (not every permutation gives rise to an automorphism, however, as Example 2 above illustrates). It was the investigation of permutations of the roots of equations that led Galois to the theory we are describing.

We have associated to each field extension K/F (equivalently, with a subfield F of K) a *group*, $\mathrm{Aut}(K/F)$, the group of automorphisms of K which fix F. One can also reverse this process and associate to each group of automorphisms a field extension.

Proposition 3. Let $H \leq \mathrm{Aut}(K)$ be a subgroup of the group of automorphisms of K. Then the collection F of elements of K fixed by all the elements of H is a subfield of K.

Proof: Let $h \in H$ and let $a, b \in F$. Then by definition $h(a) = a$, $h(b) = b$ so that $h(a \pm b) = h(a) \pm h(b) = a \pm b$, $h(ab) = h(a)h(b) = ab$ and $h(a^{-1}) = h(a)^{-1} = a^{-1}$, so that F is closed, hence a subfield of K.

Note that it is not important in this proposition that H actually be a *subgroup* of $\mathrm{Aut}(K)$ – the collection of elements of K fixed by all the elements of a *subset* of $\mathrm{Aut}(K)$ is also a subfield of K.

Definition. If H is a subgroup of the group of automorphisms of K the subfield of K fixed by all the elements of H is called the *fixed field* of H.

Proposition 4. The association of groups to fields and fields to groups defined above is inclusion reversing, namely

(1) if $F_1 \subseteq F_2 \subseteq K$ are two subfields of K then $\mathrm{Aut}(K/F_2) \leq \mathrm{Aut}(K/F_1)$, and

(2) if $H_1 \leq H_2 \leq \mathrm{Aut}(K)$ are two subgroups of automorphisms with associated fixed fields F_1 and F_2, respectively, then $F_2 \subseteq F_1$.

Proof: Any automorphism of K that fixes F_2 also fixes its subfield F_1, which gives (1). The second assertion is proved similarly.

Examples

(1) Suppose $K = \mathbb{Q}(\sqrt{2})$ as in Example 1 above. Then the fixed field of $\mathrm{Aut}(\mathbb{Q}(\sqrt{2})) = \mathrm{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) = \{1, \sigma\}$ will be the set of elements of $\mathbb{Q}(\sqrt{2})$ with

$$\sigma(a + b\sqrt{2}) = a + b\sqrt{2}$$

since everything is fixed by the identity automorphism. This is the equation

$$a - b\sqrt{2} = a + b\sqrt{2}.$$

which is equivalent to $b = 0$, so the fixed field of $\mathrm{Aut}(\mathbb{Q}(\sqrt{2})/\mathbb{Q})$ is just $\mathbb{Q}$.

(2) Suppose now that $K = \mathbb{Q}(\sqrt[3]{2})$ as in Example 2 above. In this case $\mathrm{Aut}(K) = 1$, so that every element of K is fixed, i.e. the fixed field of $\mathrm{Aut}(\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q})$ is $\mathbb{Q}(\sqrt[3]{2})$.

Given a subfield F of K, the associated group is the collection of automorphisms of K which fix F. Given a group of automorphisms of K, the associated extension is defined by taking F to be the fixed field of the automorphisms. In the first example above starting with the subfield $\mathbb{Q}$ of $\mathbb{Q}(\sqrt{2})$ one obtains the group $\{1, \sigma\}$ and starting with the group $\{1, \sigma\}$ one obtains the subfield $\mathbb{Q}$, so there is a "duality" between the two. In the second example, however, starting with the subfield $\mathbb{Q}$ of $\mathbb{Q}(\sqrt[3]{2})$ one obtains only the trivial group and starting with the trivial group one obtains the full field $\mathbb{Q}(\sqrt[3]{2})$.

An examination of the two examples suggests that for the second example there are "not enough" automorphisms to force the fixed field to be $\mathbb{Q}$ rather than the full $\mathbb{Q}(\sqrt[3]{2})$. This in turn seems to be due to the fact that the other roots of $x^3 - 2$, which are the only possible images of $\sqrt[3]{2}$ under an automorphism, are not elements of $\mathbb{Q}(\sqrt[3]{2})$. (Although even if they were we would need to check that the additional maps we could define were *automorphisms*.) We now make precise the notion of fields with "enough" automorphisms (leading to the definition of a *Galois* extension). As one might suspect even from these two examples (and we prove in the next section) these are related to splitting fields.

We first investigate the size of the automorphism group in the case of splitting fields.

Let F be a field and let E be the splitting field over F of $f(x) \in F[x]$. The main tool is Theorem 13.27 on the existence of extensions of isomorphisms, which states that any isomorphism $\varphi : F \xrightarrow{\sim} F'$ of F with F' can be extended to an isomorphism $\sigma : E \xrightarrow{\sim} E'$ between E and the splitting field E' for $f'(x) = \varphi(f(x)) \in F'[x]$.

We now show by induction on $[E : F]$ that the number of such extensions is at most $[E : F]$, with equality if $f(x)$ is separable over F. If $[E : F] = 1$ then $E = F$, $E' = F'$, $\sigma = \varphi$ and the number of extensions is 1. If $[E : F] > 1$ then $f(x)$ has at least one irreducible factor $p(x)$ of degree > 1 with corresponding irreducible factor $p'(x)$ of $f'(x)$. Let α be a fixed root of $p(x)$. If σ is any extension of φ to E, then σ restricted to the subfield $F(\alpha)$ of E is an isomorphism τ of $F(\alpha)$ with some subfield of E'. The isomorphism τ is completely determined by its action on α, i.e. by $\tau\alpha$, since α generates $F(\alpha)$ over F. Just as in Proposition 2, we see that $\tau\alpha$ must be some root

β of $p'(x)$. Then we have a diagram

$$\begin{array}{lccc} \sigma: & E & \xrightarrow{\sim} & E' \\ & | & & | \\ \tau: & F(\alpha) & \xrightarrow{\sim} & F'(\beta) \\ & | & & | \\ \varphi: & F & \xrightarrow{\sim} & F'. \end{array}$$

Conversely, for any β a root of $p'(x)$ there are extensions τ and σ giving such a diagram (this is Theorem 13.8 and Theorem 13.27). Hence to count the number of extensions σ we need only count the possible number of these diagrams.

The number of extensions of φ to an isomorphism τ is equal to the number of distinct roots β of $p'(x)$. Since the degree of $p(x)$ and $p'(x)$ are both equal to $[F(\alpha) : F]$, we see that the number of extensions of φ to a τ is at most $[F(\alpha) : F]$, with equality if the roots of $p(x)$ are distinct.

Since E is also the splitting field of $f(x)$ over $F(\alpha)$, E' is the splitting field of $f'(x)$ over $F'(\beta)$, and $[E : F(\alpha)] < [E : F]$, we may apply our induction hypothesis to these field extensions. By induction, the number of extensions of τ to σ is $\leq [E : F(\alpha)]$, with equality if $f(x)$ has distinct roots.

From $[E : F] = [E : F(\alpha)][F(\alpha) : F]$ it follows that the number of extensions of φ to σ is $\leq [E : F]$. We have equality if $p(x)$ and $f(x)$ have distinct roots, which is equivalent to $f(x)$ having distinct roots since $p(x)$ is a factor of $f(x)$, completing the proof by induction.

In the particular case when $F = F'$ and φ is the identity map we have $f(x) = f'(x)$ and $E = E'$ so the isomorphisms of E to E' restricting to φ on F are the automorphisms of E fixing F. We state this as:

Proposition 5. Let E be the splitting field over F of the polynomial $f(x) \in F[x]$. Then

$$|\mathrm{Aut}(E/F)| \leq [E : F]$$

with equality if $f(x)$ is separable over F.

Remark: While we were primarily interested in counting the automorphisms of E which fix F (which is the situation of $F = F'$, $\varphi = 1$ above), it would still have been necessary to consider the situation of more general φ (and different fields F') because of the induction step in the proof (which involves the fields $F(\alpha)$ and $F(\beta)$ for two roots of the same polynomial $p(x)$).

One can modify the proof above to show more generally that $|\mathrm{Aut}(K/F)| \leq [K : F]$ for *any* finite extension K/F (we shall prove this in the next section from a slightly different point of view). This gives us a notion of field extensions with "enough" automorphisms.

Definition. Let K/F be a finite extension. Then K is said to be *Galois* over F and K/F is a *Galois* extension if $|\mathrm{Aut}(K/F)| = [K : F]$. If K/F is Galois the group of automorphisms $\mathrm{Aut}(K/F)$ is called the *Galois group of* K/F, denoted $\mathrm{Gal}(K/F)$.

Remark: The Galois group of an extension K/F is sometimes defined to be the group of automorphisms $\mathrm{Aut}(K/F)$ for all K/F. We have chosen the definition above so that the notation $\mathrm{Gal}(K/F)$ will emphasize that the extension K/F has the maximal number of automorphisms.

Corollary 6. If K is the splitting field over F of a separable polynomial $f(x)$ then K/F is Galois.

We shall see in the next section that the converse is also true, which will completely characterize Galois extensions.

Note also that Corollary 6 implies that the splitting field of *any* polynomial over $\mathbb{Q}$ is Galois, since the splitting field of $f(x)$ is clearly the same as the splitting field of the product of the irreducible factors of $f(x)$ (i.e. the polynomial obtained by removing multiple factors), which is separable (Corollary 13.34).

Definition. If $f(x)$ is a separable polynomial over F, then the *Galois group of $f(x)$ over F* is the Galois group of the splitting field of $f(x)$ over F.

Examples

(1) The extension $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ is Galois with Galois group $\mathrm{Gal}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) = \{1, \sigma\} \cong \mathbb{Z}/2\mathbb{Z}$ where σ is the automorphism

$$\begin{aligned} \sigma : \mathbb{Q}(\sqrt{2}) &\xrightarrow{\sim} \mathbb{Q}(\sqrt{2}) \\ a + b\sqrt{2} &\mapsto a - b\sqrt{2}. \end{aligned}$$

(2) More generally, any quadratic extension K of any field F of characteristic different from 2 is Galois. This follows from the discussion of quadratic extensions following Corollary 13.13, which shows that any extension K of degree 2 of F (where the characteristic of F is not 2) is of the form $F(\sqrt{D})$ for some D hence is the splitting field of $x^2 - D$ (since if $\sqrt{D} \in K$ then also $-\sqrt{D} \in K$).

(3) The extension $\mathbb{Q}(\sqrt[3]{2})/\mathbb{Q}$ is not Galois since its group of automorphisms is only of order 1.

(4) The extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ is Galois over $\mathbb{Q}$ since it is the splitting field of the polynomial $(x^2-2)(x^2-3)$. Any automorphism σ is completely determined by its action on the generators $\sqrt{2}$ and $\sqrt{3}$, which must be mapped to $\pm\sqrt{2}$ and $\pm\sqrt{3}$, respectively. Hence the only possibilities for automorphisms are the maps

$$\begin{cases} \sqrt{2} \mapsto \sqrt{2} \\ \sqrt{3} \mapsto \sqrt{3} \end{cases} \quad \begin{cases} \sqrt{2} \mapsto -\sqrt{2} \\ \sqrt{3} \mapsto \sqrt{3} \end{cases} \quad \begin{cases} \sqrt{2} \mapsto \sqrt{2} \\ \sqrt{3} \mapsto -\sqrt{3} \end{cases} \quad \begin{cases} \sqrt{2} \mapsto -\sqrt{2} \\ \sqrt{3} \mapsto -\sqrt{3} \end{cases}.$$

Since the Galois group is of order 4, *all* these elements are in fact automorphisms of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$.

Define the automorphisms σ and τ by

$$\sigma : \begin{cases} \sqrt{2} \mapsto -\sqrt{2} \\ \sqrt{3} \mapsto \sqrt{3} \end{cases} \qquad \tau : \begin{cases} \sqrt{2} \mapsto \sqrt{2} \\ \sqrt{3} \mapsto -\sqrt{3} \end{cases}$$

or, more explicitly, by

$$\sigma : a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mapsto a - b\sqrt{2} + c\sqrt{3} - d\sqrt{6}$$
$$\tau : a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mapsto a + b\sqrt{2} - c\sqrt{3} - d\sqrt{6}$$

(since, for example,

$$\sigma(\sqrt{6}) = \sigma(\sqrt{2}\sqrt{3}) = \sigma(\sqrt{2})\sigma(\sqrt{3}) = (-\sqrt{2})(\sqrt{3}) = -\sqrt{6}).$$

Then $\sigma^2(\sqrt{2}) = \sigma(\sigma\sqrt{2}) = \sigma(-\sqrt{2}) = \sqrt{2}$ and clearly $\sigma^2(\sqrt{3}) = \sqrt{3}$. Hence $\sigma^2 = 1$ is the identity automorphism. Similarly, $\tau^2 = 1$. The automorphism $\sigma\tau$ can be easily computed:

$$\sigma\tau(\sqrt{2}) = \sigma(\tau(\sqrt{2})) = \sigma(\sqrt{2}) = -\sqrt{2}$$

and

$$\sigma\tau(\sqrt{3}) = \sigma(\tau(\sqrt{3})) = \sigma(-\sqrt{3}) = -\sqrt{3}$$

so that $\sigma\tau$ is the remaining nontrivial automorphism in the Galois group. Since this automorphism also evidently has order 2 in the Galois group, we have

$$\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}) = \{1, \sigma, \tau, \sigma\tau\}$$

i.e. the Galois group is isomorphic to the Klein 4-group.

Associated to each subgroup of $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q})$ is the corresponding fixed subfield of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. For example, the subfield corresponding to $\{1, \sigma\tau\}$ is the set of elements fixed by the map

$$\sigma\tau : a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mapsto a - b\sqrt{2} - c\sqrt{3} + d\sqrt{6}$$

which is the set of elements $a + d\sqrt{6}$ i.e. the field $\mathbb{Q}(\sqrt{6})$. One can similarly determine the fixed fields for the other subgroups of the Galois group:

subgroup	fixed field
$\{1\}$	$\mathbb{Q}(\sqrt{2}, \sqrt{3})$
$\{1, \sigma\}$	$\mathbb{Q}(\sqrt{3})$
$\{1, \sigma\tau\}$	$\mathbb{Q}(\sqrt{6})$
$\{1, \tau\}$	$\mathbb{Q}(\sqrt{2})$
$\{1, \sigma, \tau, \sigma\tau\}$	$\mathbb{Q}$.

(5) The splitting field of $x^3 - 2$ over $\mathbb{Q}$ is Galois of degree 6. The roots of this equation are $\sqrt[3]{2}$, $\rho\sqrt[3]{2}$, $\rho^2\sqrt[3]{2}$ where $\rho = \zeta_3 = \dfrac{-1 + \sqrt{-3}}{2}$ is a primitive cube root of unity. Hence the splitting field can be written $\mathbb{Q}(\sqrt[3]{2}, \rho\sqrt[3]{2})$. Any automorphism maps each of these two elements to one of the roots of $x^3 - 2$, giving 9 possibilities, but since the Galois group has order 6 not every such map is an automorphism of the field.

To determine the Galois group we use a more convenient set of generators, namely $\sqrt[3]{2}$ and ρ. Then any automorphism σ maps $\sqrt[3]{2}$ to one of $\sqrt[3]{2}$, $\rho\sqrt[3]{2}$, $\rho^2\sqrt[3]{2}$, and maps ρ to ρ or $\rho^2 = \dfrac{-1 - \sqrt{-3}}{2}$ since these are the roots of the cyclotomic

polynomial $\Phi_3(x) = x^2 + x + 1$. Since σ is completely determined by its action on these two elements this gives only 6 possibilities and so each of these possibilities is actually an automorphism. To give these automorphisms explicitly, let σ and τ be the automorphisms defined by

$$\sigma : \begin{cases} \sqrt[3]{2} \mapsto \rho\sqrt[3]{2} \\ \rho \mapsto \rho \end{cases} \qquad \tau : \begin{cases} \sqrt[3]{2} \mapsto \sqrt[3]{2} \\ \rho \mapsto \rho^2 = -1 - \rho. \end{cases}$$

As before, these can be given explicitly on the elements of $\mathbb{Q}(\sqrt[3]{2}, \rho)$, which are linear combinations of the basis $\{1, \sqrt[3]{2}, (\sqrt[3]{2})^2, \rho, \rho\sqrt[3]{2}, \rho(\sqrt[3]{2})^2\}$. For example

$$\begin{aligned} \sigma(\rho\sqrt[3]{2}) &= (\rho)(\rho\sqrt[3]{2}) = \rho^2\sqrt[3]{2} = (-1-\rho)\sqrt[3]{2} \\ &= -\sqrt[3]{2} - \rho\sqrt[3]{2} \end{aligned}$$

and we may similarly determine the action of σ on the other basis elements. This gives

$$\begin{aligned} \sigma : \quad a + b\sqrt[3]{2} + c\sqrt[3]{4} + d\rho + e\rho\sqrt[3]{2} + f\rho\sqrt[3]{4} \quad &\mapsto \\ a - e\sqrt[3]{2} + (f-c)\sqrt[3]{4} + d\rho + (b-e)\rho\sqrt[3]{2} - c\rho\sqrt[3]{4}. \end{aligned} \tag{14.1}$$

The other elements of the Galois group are

$$1 : \begin{cases} \sqrt[3]{2} \mapsto \sqrt[3]{2} \\ \rho \mapsto \rho \end{cases} \qquad \sigma^2 : \begin{cases} \sqrt[3]{2} \mapsto \rho^2\sqrt[3]{2} \\ \rho \mapsto \rho \end{cases}$$

$$\tau\sigma : \begin{cases} \sqrt[3]{2} \mapsto \rho^2\sqrt[3]{2} \\ \rho \mapsto \rho^2 \end{cases} \qquad \tau\sigma^2 : \begin{cases} \sqrt[3]{2} \mapsto \rho\sqrt[3]{2} \\ \rho \mapsto \rho^2. \end{cases}$$

Computing $\sigma\tau$ we have

$$\sigma\tau : \begin{cases} \sqrt[3]{2} \overset{\tau}{\mapsto} \sqrt[3]{2} \overset{\sigma}{\mapsto} \rho\sqrt[3]{2} \\ \rho \overset{\tau}{\mapsto} \rho^2 \overset{\sigma}{\mapsto} \rho^2 \end{cases}$$

i.e.

$$\sigma\tau : \begin{cases} \sqrt[3]{2} \mapsto \rho\sqrt[3]{2} \\ \rho \mapsto \rho^2 \end{cases}$$

so that $\sigma\tau = \tau\sigma^2$. Similarly one computes that $\sigma^3 = \tau^2 = 1$. Hence

$$\mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2}, \zeta_3)/\mathbb{Q}) = \langle \sigma, \tau \rangle \cong S_3$$

is the symmetric group on 3 letters. Alternatively (and less computationally), since $G = \mathrm{Gal}(\mathbb{Q}(\sqrt[3]{2}, \zeta_3)/\mathbb{Q})$ acts as permutations of the 3 roots of $x^3 - 2$, G is a subgroup of S_3, hence must be S_3 since it is of order 6. The computations above explicitly identify the automorphisms in G and give an explicit isomorphism of G with S_3.

As in the previous example we can determined the fixed fields for any of the subgroups of the Galois group. For example, consider the fixed field of the subgroup

$\{1, \sigma, \sigma^2\}$ generated by σ. These are just the elements fixed by σ (given explicitly in equation (1)) since if an element is fixed by σ then it is also fixed by σ^2. (In general, the fixed field of some subgroup is the field fixed by a set of generators for the subgroup.) The elements fixed by σ are those with

$$a = a \quad b = -e \quad c = f - c \quad d = d \quad e = b - e \quad f = -c$$

which is equivalent to $b = c = f = e = 0$. Hence the fixed field of $\{1, \sigma, \sigma^2\}$ is the field $\mathbb{Q}(\rho)$.

Remark: This example shows that some care must be exercised in determining Galois groups from the actions on generators. As mentioned, not every map taking $\sqrt[3]{2}$ and $\rho\sqrt[3]{2}$ to roots of $x^3 - 2$ gives rise to an automorphism of the field (for example, the map

$$\begin{aligned} \sqrt[3]{2} &\mapsto \rho\sqrt[3]{2} \\ \rho\sqrt[3]{2} &\mapsto \rho\sqrt[3]{2} \end{aligned}$$

clearly cannot be an automorphism since it is evidently not an injection). The point is that there may be (sometimes very subtle) algebraic relations among the generators and these relations must be respected by an automorphism. For example, the quotient of the generators here is ρ, which is mapped to 1 and not to a root of the minimal polynomial for ρ. Put another way, the quotient of these generators satisfies a quadratic equation and this map does not respect that property.

For another (less trivial) example, compare with the discussion of the splitting field of $x^8 - 2$ in Section 2.

(6) As in Example 3, the field $\mathbb{Q}(\sqrt[4]{2})$ is not Galois over $\mathbb{Q}$ since any automorphism is determined by where it sends $\sqrt[4]{2}$ and of the four possibilities $\{\pm\sqrt[4]{2}, \pm i\sqrt[4]{2}\}$, only two are elements of the field (the two real roots).

Note that we have

$$\overbrace{\underbrace{\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})}_{2} \underbrace{\subseteq \mathbb{Q}(\sqrt[4]{2})}_{2}}^{4}$$

where $\mathbb{Q}(\sqrt{2})/\mathbb{Q}$ and $\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}(\sqrt{2})$ are both Galois extensions by Example 2 since both are quadratic extensions. This shows that a Galois extension of a Galois extension is not necessarily Galois.

(7) The extension of finite fields $\mathbb{F}_{p^n}/\mathbb{F}_p$ constructed after Proposition 13.37 is Galois by Corollary 6 since $\mathbb{F}_{p^n}$ is the splitting field over $\mathbb{F}_p$ of the separable polynomial $x^{p^n} - x$. It follows that the group of automorphisms for this extension is of order n. The injective homomorphism

$$\begin{aligned} \sigma : \mathbb{F}_{p^n} &\to \mathbb{F}_{p^n} \\ \alpha &\mapsto \alpha^p \end{aligned}$$

of Proposition 13.35 is surjective in this case since $\mathbb{F}_{p^n}$ is finite, hence is an isomorphism. This gives an automorphism of $\mathbb{F}_{p^n}$, called the *Frobenius* automorphism, which we shall denote by σ_p. Iterating σ_p we have $\sigma_p{}^2(\alpha) = \sigma_p(\sigma_p(\alpha)) = (\alpha^p)^p = \alpha^{p^2}$. Similarly we have

$$\sigma_p{}^i(\alpha) = \alpha^{p^i} \qquad i = 0, 1, 2, \ldots$$

Since $\alpha^{p^n} = \alpha$, we see that $\sigma_p{}^{p^n} = 1$ is the identity automorphism. No lower power of σ_p can be the identity, since this would imply $\alpha^{p^i} = \alpha$ for all $\alpha \in \mathbb{F}_{p^n}$ for some $i < n$, which is impossible since there are only p^i roots of this equation. It follows that σ_p is of order n in the Galois group, which means that $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$ is *cyclic* of order n, with the Frobenius automorphism σ_p as generator.

(8) The inseparable extension $\mathbb{F}_2(x)$ over $\mathbb{F}_2(t)$ where $x^2 - t = 0$ considered in Section 13.5 is not Galois. Any automorphism of this degree 2 extension is determined by its action on x, which must be sent to a root of the equation $x^2 - t$. We have already seen that there is only one root of this equation (with multiplicity 2) since we are in a field of characteristic 2. Hence the extension has only the trivial automorphism. Note that $\mathbb{F}_2(x)$ is the splitting field for $x^2 - t$ over $\mathbb{F}_2(t)$, so this example shows the separability condition in Corollary 6 is necessary.

EXERCISES

1. **(a)** Show that if the field K is generated over F by the elements $\alpha_1, \ldots, \alpha_n$ then an automorphism σ of K fixing F is uniquely determined by $\sigma(\alpha_1), \ldots, \sigma(\alpha_n)$. In particular show that an automorphism fixes K if and only if it fixes a set of generators for K.
   (b) Let $G \leq \mathrm{Gal}(K/F)$ be a subgroup of the Galois group of the extension K/F and suppose $\sigma_1, \ldots, \sigma_k$ are generators for G. Show that the subfield E/F is fixed by G if and only if it is fixed by the generators $\sigma_1, \ldots, \sigma_k$.
2. Let τ be the map $\tau : \mathbb{C} \to \mathbb{C}$ defined by $\tau(a + bi) = a - bi$ (*complex conjugation*). Prove that τ is an automorphism of $\mathbb{C}$.
3. Determine the fixed field of complex conjugation on $\mathbb{C}$.
4. Prove that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$ are not isomorphic.
5. Determine the automorphisms of the extension $\mathbb{Q}(\sqrt[4]{2})/\mathbb{Q}(\sqrt{2})$ explicitly.
6. Let F be a field.
   (a) Show that the mapping $\varphi : F[x] \to F[x]$ defined by $\varphi(f(x)) = f(ax + b)$ for fixed $a, b \in F$, $a \neq 0$ is an automorphism of $F[x]$ which is the identity on F.
   (b) Conversely, let φ be an automorphism of $F[x]$ which is the identity on F. Prove that there exist $a, b \in F$ with $a \neq 0$ such that $\varphi(f(x)) = f(ax + b)$ as in (a).
7. This exercise determines $\mathrm{Aut}(\mathbb{R}/\mathbb{Q})$.
   (a) Prove that any $\sigma \in \mathrm{Aut}(\mathbb{R}/\mathbb{Q})$ takes squares to squares and takes positive reals to positive reals. Conclude that $a < b$ implies $\sigma a < \sigma b$ for every $a, b \in \mathbb{R}$.
   (b) Prove that $-\frac{1}{m} < a - b < \frac{1}{m}$ implies $-\frac{1}{m} < \sigma a - \sigma b < \frac{1}{m}$. Conclude that σ is a continuous map on $\mathbb{R}$.
   (c) Prove that any continuous map on $\mathbb{R}$ which is the identity on $\mathbb{Q}$ is the identity map, hence $\mathrm{Aut}(\mathbb{R}/\mathbb{Q}) = 1$.
8. Prove that the automorphisms of the rational function field $F(x)$ which fix F are precisely the *fractional linear transformations* determined by $x \mapsto \dfrac{ax + b}{cx + d}$ for $a, b, c, d \in F$, $ad - bc \neq 0$ (so $f(x) \in F(x)$ maps to $f(\dfrac{ax + b}{cx + d})$) (cf. Exercise 18 of Section 13.2).
9. Determine the fixed field of the automorphism $x \mapsto x + 1$ of $F(x)$.
10. Let K be an extension of the field F. Let $\varphi : K \to K'$ be an isomorphism of K with a field K' which maps F to the subfield F' of K'. Prove that the map $\sigma \mapsto \varphi\sigma\varphi^{-1}$ defines a group isomorphism $\mathrm{Aut}(K/F) \xrightarrow{\sim} \mathrm{Aut}(K'/F')$.

14.2 THE FUNDAMENTAL THEOREM OF GALOIS THEORY

In the Galois extension $\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q})$ considered in the previous section, there was a strong similarity between the diagram of subgroups of the Galois group :

$$\begin{tikzcd}
 & \{1\} \arrow[dl, dash, "2"'] \arrow[d, dash, "2"] \arrow[dr, dash, "2"] & \\
\{1, \sigma\tau\} \arrow[dr, dash, "2"'] & \{1, \sigma\tau\} \arrow[d, dash, "2"] & \{1, \tau\} \arrow[dl, dash, "2"] \\
 & \{1, \sigma, \tau, \sigma\tau\} &
\end{tikzcd}$$

and the diagram of corresponding fixed fields

$$\begin{tikzcd}
 & \mathbb{Q}(\sqrt{2}, \sqrt{3}) \arrow[dl, dash, "2"'] \arrow[d, dash, "2"] \arrow[dr, dash, "2"] & \\
\mathbb{Q}(\sqrt{2}) \arrow[dr, dash, "2"'] & \mathbb{Q}(\sqrt{6}) \arrow[d, dash, "2"] & \mathbb{Q}(\sqrt{3}) \arrow[dl, dash, "2"] \\
 & \mathbb{Q} &
\end{tikzcd}$$

(we have inverted the lattice of subgroups because of the inclusion-reversing nature of the correspondence).

Note that this is also the diagram of *all* known subfields of the extension and that in this case each of the subfields is also a Galois extension of $\mathbb{Q}$.

In a similar way there is a strong similarity between the diagram of subgroups of the Galois group and known subfields for the splitting field of $x^3 - 2$:

$$\begin{tikzcd}
 & 1 \arrow[dl, dash] \arrow[dr, dash] \arrow[drr, dash] \arrow[drrr, dash] & & & \\
\langle\sigma\rangle \arrow[dr, dash] & & \langle\tau\rangle \arrow[dl, dash] & \langle\tau\sigma\rangle \arrow[dll, dash] & \langle\tau\sigma^2\rangle \arrow[dlll, dash] \\
 & \langle\sigma, \tau\rangle & & &
\end{tikzcd}$$

$$\begin{tikzcd}
 & \mathbb{Q}(\sqrt[3]{2}, \rho) \arrow[dl, dash] \arrow[dr, dash] \arrow[drr, dash] \arrow[drrr, dash] & & & \\
\mathbb{Q}(\rho) \arrow[dr, dash] & & \mathbb{Q}(\sqrt[3]{2}) \arrow[dl, dash] & \mathbb{Q}(\rho\sqrt[3]{2}) \arrow[dll, dash] & \mathbb{Q}(\rho^2\sqrt[3]{2}) \arrow[dlll, dash] \\
 & \mathbb{Q} & & &
\end{tikzcd}$$

where the subfields in the second diagram are precisely the fixed fields of the subgroups in the first diagram.

Note in this pair of diagrams that only the subgroup $\langle\sigma\rangle$ generated by σ is normal in S_3 and that the subfield $\mathbb{Q}(\rho)$ is the only subfield Galois over $\mathbb{Q}$.

The Fundamental Theorem of Galois Theory states that the relations observed in the two examples above are not coincidental and hold for any Galois extension. Before proving this we first develop some preliminary results on *group characters*, of which field automorphisms give particular examples.

Definition. A *character*[1] χ of a group G with values in a field L is a homomorphism from G to the multiplicative group of L:

$$\chi : G \to L^\times$$

i.e. $\chi(g_1g_2) = \chi(g_1)\chi(g_2)$ for all $g_1, g_2 \in G$ and $\chi(g)$ is a nonzero element of L for all $g \in G$.

Definition. The characters $\chi_1, \chi_2, ..., \chi_n$ of G are said to be *linearly independent* over L if they are linearly independent as functions on G, i.e. if there is no nontrivial relation

$$a_1\chi_1 + a_2\chi_2 + ... + a_n\chi_n = 0 \qquad (a_1, ..., a_n \in L \text{ not all } 0) \tag{14.2}$$

as a function on G (that is, $a_1\chi_1(g) + a_2\chi_2(g) + ... + a_n\chi_n(g) = 0$ for all $g \in G$).

Theorem 7. (Linear Independence of Characters) If $\chi_1, \chi_2, ..., \chi_n$ are distinct characters of G with values in L then they are linearly independent over L.

Proof: Suppose the characters were linearly dependent. Among all the linear dependence relations (2) above, choose one with the minimal number m of nonzero coefficients a_i. We may suppose (by renumbering, if necessary) that the m nonzero coefficients are $a_1, a_2, ..., a_m$:

$$a_1\chi_1 + a_2\chi_2 + ... + a_m\chi_m = 0.$$

Then for any $g \in G$ we have

$$a_1\chi_1(g) + a_2\chi_2(g) + ... + a_m\chi_m(g) = 0. \tag{14.3}$$

Let g_0 be an element with $\chi_1(g_0) \neq \chi_m(g_0)$ (which exists, since $\chi_1 \neq \chi_m$). Since (3) holds for every element of G, in particular we have

$$a_1\chi_1(g_0g) + a_2\chi_2(g_0g) + ... + a_m\chi_m(g_0g) = 0$$

i.e.

$$a_1\chi_1(g_0)\chi_1(g) + a_2\chi_2(g_0)\chi_2(g) + ... + a_m\chi_m(g_0)\chi_m(g) = 0. \tag{14.4}$$

Multiplying equation (3) by $\chi_m(g_0)$ and subtracting from equation (4) we obtain

$$[\chi_m(g_0) - \chi_1(g_0)]a_1\chi_1(g) + [\chi_m(g_0) - \chi_2(g_0)]a_2\chi_2(g) + ...$$
$$+[\chi_m(g_0) - \chi_{m-1}(g_0)]a_{m-1}\chi_{m-1}(g) = 0,$$

[1] This is the definition of a *linear* character. More general characters will be studied in Chapter 15.

which holds for all $g \in G$. But the first coefficient is nonzero and this is a relation with fewer nonzero coefficients, a contradiction.

Consider now an injective homomorphism σ of a field K into a field L, called an *embedding* of K into L. Then in particular σ is a homomorphism of the multiplicative group $G = K^\times$ into the multiplicative group $L^\times$, so σ may be viewed as a character of $K^\times$ with values in L. Note also that this character contains all of the useful information about the values of σ viewed simply as a *function* on K, since the only point of K not considered in $K^\times$ is 0, and we know σ maps 0 to 0.

Corollary 8. If $\sigma_1, \sigma_2, ..., \sigma_n$ are distinct embeddings of a field K into a field L, then they are linearly independent as functions on K. In particular distinct automorphisms of a field K are linearly independent as functions on K.

We now use Corollary 8 to prove the fundamental relation between the orders of subgroups of the automorphism group of a field K and the degrees of the extensions over their fixed fields.

Theorem 9. Let $G = \{\sigma_1 = 1, \sigma_2, ..., \sigma_n\}$ be a subgroup of automorphisms of a field K and let F be the fixed field. Then

$$[K : F] = n = |G|.$$

Proof: Suppose first that $n > [K : F]$.

Let $\omega_1, \omega_2, ..., \omega_m$ be a basis for K over F ($m = [K : F]$). Then the system

$$\begin{gathered}\sigma_1(\omega_1)x_1 + \sigma_2(\omega_1)x_2 + ... + \sigma_n(\omega_1)x_n = 0 \\ \vdots \\ \sigma_1(\omega_m)x_1 + \sigma_2(\omega_m)x_2 + ... + \sigma_n(\omega_m)x_n = 0\end{gathered}$$

of m equations in n unknowns $x_1, x_2, ..., x_n$ has a nontrivial solution $\beta_1, \beta_2, ..., \beta_n$ in K since by assumption there are more unknowns than equations.

Let $a_1, a_2, ..., a_m$ be m arbitrary elements of F. The field F is by definition fixed by $\sigma_1, ..., \sigma_n$ so each of these elements is fixed by every σ_i, i.e. $\sigma_i(a_j) = a_j$, $i = 1, 2, ..., n$, $j = 1, 2, ..., m$. Multiplying the first equation above by a_1, the second by a_2,..., the last by a_m then gives the system of equations

$$\begin{gathered}\sigma_1(a_1\omega_1)\beta_1 + \sigma_2(a_1\omega_1)\beta_2 + ... + \sigma_n(a_1\omega_1)\beta_n = 0 \\ \vdots \\ \sigma_1(a_m\omega_m)\beta_1 + \sigma_2(a_m\omega_m)\beta_2 + ... + \sigma_n(a_m\omega_m)\beta_n = 0.\end{gathered}$$

Adding these equations we see that there are elements $\beta_1, ..., \beta_n$ in K, not all 0, satisfying

$$\sigma_1(a_1\omega_1 + a_2\omega_2 + ... + a_m\omega_m)\beta_1 + ... + \sigma_n(a_1\omega_1 + a_2\omega_2 + ... + a_m\omega_m)\beta_n = 0$$

for all choices of $a_1, ..., a_m$ in F. Since $\omega_1, ..., \omega_m$ is an F-basis for K, every $\alpha \in K$ is of the form $a_1\omega_1 + a_2\omega_2 + ... + a_m\omega_m$, so the previous equation means

$$\sigma_1(\alpha)\beta_1 + ... + \sigma_n(\alpha)\beta_n = 0$$

for all $\alpha \in K$. But this means the distinct automorphisms $\sigma_1, ..., \sigma_n$ are linearly dependent over K, contradicting Corollary 8.

We have proved $n \leq [K : F]$. Note that we have so far not used the fact that $\sigma_1, \sigma_2, ..., \sigma_n$ are the elements of a *group*.

Suppose now that $n < [K : F]$.

Then there are more than n F-linearly independent elements of K, say $\alpha_1, \alpha_2, ..., \alpha_{n+1}$. The system

$$\begin{aligned} \sigma_1(\alpha_1)x_1 + \sigma_1(\alpha_2)x_2 + ... + \sigma_1(\alpha_{n+1})x_{n+1} &= 0 \\ &\vdots \\ \sigma_n(\alpha_1)x_1 + \sigma_n(\alpha_2)x_2 + ... + \sigma_n(\alpha_{n+1})x_{n+1} &= 0 \end{aligned} \tag{14.5}$$

of n equations in $n + 1$ unknowns $x_1, ..., x_{n+1}$ has a solution $\beta_1, ..., \beta_{n+1}$ in K where not all the β_i, $i = 1, 2, ..., n + 1$ are 0. If all the elements of the solution $\beta_1, ..., \beta_{n+1}$ were elements of F then the first equation (recall $\sigma_1 = 1$ is the identity automorphism) would contradict the linear independence over F of $\alpha_1, \alpha_2, ..., \alpha_{n+1}$. Hence at least one β_i, $i = 1, 2, ..., n + 1$, is not an element of F.

Among all the nontrivial solutions $(\beta_1, ..., \beta_{n+1})$ of the system (5) choose one with the minimal number r of nonzero β_i. By renumbering if necessary we may assume $\beta_1, ..., \beta_r$ are nonzero. Dividing the equations by β_r we may also assume $\beta_r = 1$. We have already seen that at least one of $\beta_1, ..., \beta_{r-1}, 1$ is not an element of F (which shows in particular that $r > 1$), say $\beta_1 \notin F$. Then our system of equations reads

$$\begin{aligned} \sigma_1(\alpha_1)\beta_1 + ... + \sigma_1(\alpha_{r-1})\beta_{r-1} + \sigma_1(\alpha_r) &= 0 \\ &\vdots \\ \sigma_n(\alpha_1)\beta_1 + ... + \sigma_n(\alpha_{r-1})\beta_{r-1} + \sigma_n(\alpha_r) &= 0 \end{aligned} \tag{14.6}$$

or more briefly

$$\sigma_i(\alpha_1)\beta_1 + ... + \sigma_i(\alpha_{r-1})\beta_{r-1} + \sigma_i(\alpha_r) = 0 \qquad i = 1, 2, ..., n. \tag{14.7}$$

Since $\beta_1 \notin F$, there is some automorphism σ_{k_0} ($k_0 \in \{1, 2, ..., n\}$) with

$$\sigma_{k_0}\beta_1 \neq \beta_1.$$

If we apply the automorphism σ_{k_0} to the equations in (6), we obtain the system of equations

$$\sigma_{k_0}\sigma_j(\alpha_1)\sigma_{k_0}(\beta_1) + ... + \sigma_{k_0}\sigma_j(\alpha_{r-1})\sigma_{k_0}(\beta_{r-1}) + \sigma_{k_0}\sigma_j(\alpha_r) = 0 \tag{14.8}$$

for $j = 1, 2, ..., n$. But the elements

$$\sigma_{k_0}\sigma_1, \sigma_{k_0}\sigma_2, ..., \sigma_{k_0}\sigma_n$$

are the same as the elements

$$\sigma_1, \sigma_2, ..., \sigma_n$$

in some order since these elements form a *group*. In other words, if we define the index i by $\sigma_{k_0}\sigma_j = \sigma_i$ then i and j both run over the set $\{1, 2, ..., n\}$. Hence the equations in (8) can be written

$$\sigma_i(\alpha_1)\sigma_{k_0}(\beta_1) + ... + \sigma_i(\alpha_{r-1})\sigma_{k_0}(\beta_{r-1}) + \sigma_i(\alpha_r) = 0. \tag{14.8'}$$

If we now subtract the equations in (8′) from those in (7) we obtain the system

$$\sigma_i(\alpha_1)[\beta_1 - \sigma_{k_0}(\beta_1)] + \ldots + \sigma_i(\alpha_{r-1})[\beta_{r-1} - \sigma_{k_0}(\beta_{r-1})] = 0$$

for $i = 1, 2, \ldots, n$. But this is a solution to the system of equations (5) with

$$x_1 = \beta_1 - \sigma_{k_0}(\beta_1) \neq 0$$

(by the choice of k_0), hence is nontrivial and has fewer than r nonzero x_i. This is a contradiction and completes the proof.

Our first use of this result is to prove that the inequality of Proposition 5 holds for any finite extension K/F.

Corollary 10. Let K/F be any finite extension. Then

$$|\mathrm{Aut}(K/F)| \leq [K : F]$$

with equality if and only if F is the fixed field of $\mathrm{Aut}(K/F)$. Put another way, K/F is Galois if and only if F is the fixed field of $\mathrm{Aut}(K/F)$.

Proof: Let F_1 be the fixed field of $\mathrm{Aut}(K/F)$, so that

$$F \subseteq F_1 \subseteq K.$$

By Theorem 9, $[K : F_1] = |\mathrm{Aut}(K/F)|$. Hence $[K : F] = |\mathrm{Aut}(K/F)|[F_1 : F]$, which proves the corollary.

Corollary 11. Let G be a finite subgroup of automorphisms of a field K and let F be the fixed field. Then every automorphism of K fixing F is contained in G, i.e. $\mathrm{Aut}(K/F) = G$, so that K/F is Galois, with Galois group G.

Proof: By definition F is fixed by all the elements of G so we have $G \leq \mathrm{Aut}(K/F)$ (and the question is whether there are any automorphisms of K fixing F not in G i.e. whether this containment is proper). Hence $|G| \leq |\mathrm{Aut}(K/F)|$. By the theorem we have $|G| = [K : F]$ and by the previous corollary $|\mathrm{Aut}(K/F)| \leq [K : F]$. This gives

$$[K : F] = |G| \leq |\mathrm{Aut}(K/F)| \leq [K : F]$$

and it follows that we must have equalities throughout, proving the corollary.

Corollary 12. If $G_1 \neq G_2$ are distinct finite subgroups of automorphisms of a field K then their fixed fields are also distinct.

Proof: Suppose F_1 is the fixed field of G_1 and F_2 is the fixed field of G_2. If $F_1 = F_2$ then by definition F_1 is fixed by G_2. By the previous corollary any automorphism fixing F_1 is contained in G_1, hence $G_2 \leq G_1$. Similarly $G_1 \leq G_2$ and so $G_1 = G_2$.

By the corollaries above we see that taking the fixed fields for distinct finite subgroups of $\mathrm{Aut}(K)$ gives distinct subfields of K over which K is Galois. Further, the degrees of the extensions are given by the orders of the subgroups. We saw this ex-

plicitly for the fields $K = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ and $K = \mathbb{Q}(\sqrt[3]{2}, \rho)$ above. A portion of the Fundamental Theorem states that these are *all* the subfields of K.

The next result provides the converse of Proposition 5 and characterizes Galois extensions.

Theorem 13. The extension K/F is Galois if and only if K is the splitting field of some separable polynomial over F. Furthermore, if this is the case then every irreducible polynomial with coefficients in F which has a root in K is separable and has all its roots in K (so in particular K/F is a separable extension).

Proof: Proposition 5 proves that the splitting field of a separable polynomial is Galois.

We now show that if K/F is Galois then every irreducible polynomial $p(x)$ in $F[x]$ having a root in K splits completely in K. Set $G = \mathrm{Gal}(K/F)$. Let $\alpha \in K$ be a root of $p(x)$ and consider the elements

$$\alpha, \sigma_2(\alpha), ..., \sigma_n(\alpha) \in K \tag{14.9}$$

where $\{1, \sigma_2, ..., \sigma_n\}$ are the elements of $\mathrm{Gal}(K/F)$. Let

$$\alpha, \alpha_2, \alpha_3, ..., \alpha_r$$

denote the *distinct* elements in (9). If $\tau \in G$ then since G is a group the elements $\{\tau, \tau\sigma_2, ..., \tau\sigma_n\}$ are the same as the elements $\{1, \sigma_2, ..., \sigma_n\}$ in some order. It follows that applying $\tau \in G$ to the elements in (9) simply permutes them, so in particular applying τ to $\alpha, \alpha_2, \alpha_3, ..., \alpha_r$ also permutes these elements. The polynomial

$$f(x) = (x - \alpha)(x - \alpha_2)...(x - \alpha_r)$$

therefore has coefficients which are fixed by all the elements of G since the elements of G simply permute the factors. Hence the coefficients lie in the fixed field of G, which by Corollary 10 is the field F. Hence $f(x) \in F[x]$.

Since $p(x)$ is irreducible and has α as a root, $p(x)$ is the minimal polynomial for α over F, hence divides any polynomial with coefficients in F having α as a root (this is Proposition 13.9). It follows that $p(x)$ divides $f(x)$ in $F[x]$ and since $f(x)$ obviously divides $p(x)$ in $K[x]$ by Proposition 2, we have

$$p(x) = f(x).$$

In particular, this shows that $p(x)$ is separable and that all its roots lie in K (in fact they are among the elements $\alpha, \sigma_2\alpha, ..., \sigma_n\alpha$). This proves the last statement of the theorem.

To complete the proof, suppose K/F is Galois and let $\omega_1, \omega_2, ..., \omega_n$ be a basis for K/F. Let $p_i(x)$ be the minimal polynomial for ω_i over F, $i = 1, 2, ..., n$. Then by what we have just proved, $p_i(x)$ is separable and has all its roots in K. Let $g(x)$ be the polynomial obtained by removing any multiple factors in the product $p_1(x)p_2(x)...p_n(x)$ (the "square free part"). Then the splitting field of the two polynomials is the same and this field is K (all the roots lie in K, so K contains the splitting field, but $\omega_1, \omega_2, ..., \omega_n$ are among the roots, so the splitting field contains K). Hence K is the splitting field of the separable polynomial $g(x)$.

Definition. Let K/F be a Galois extension. If $\alpha \in K$ the elements $\sigma\alpha$ for $\sigma \in \mathrm{Gal}(K/F)$ are called the *conjugates* (or *Galois conjugates*) of α over F.

The proof of the theorem shows that in a Galois extension K/F the other roots of the minimal polynomial over F of any element $\alpha \in K$ are precisely the distinct conjugates of α under the Galois group of K/F.

The second statement in this theorem also shows that K is not Galois over F if we can find even one irreducible polynomial over F having a root in K but not having *all* its roots in K. This justifies in a very strong sense the intuition from earlier examples that Galois extensions are extensions with "enough" distinct roots of irreducible polynomials (namely, if it contains one root then it contains all the roots).

Finally, notice that we now have 4 characterizations of Galois extensions K/F:

(1) splitting fields of separable polynomials over F

(2) fields where F is precisely the set of elements fixed by $\mathrm{Aut}(K/F)$ (in general, the fixed field may be larger than F)

(3) fields with $[K : F] = |\mathrm{Aut}(K/F)|$ (the original definition)

(4) finite, normal and separable extensions.

Theorem 14. (Fundamental Theorem of Galois Theory) Let K/F be a Galois extension and set $G = \mathrm{Gal}(K/F)$. Then there is a bijection

$$\left\{ \begin{array}{c} \text{subfields } E \\ \text{of } K \\ \text{containing } F \end{array} \begin{array}{c} K \\ | \\ E \\ | \\ F \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{subgroups } H \\ \text{of } G \end{array} \begin{array}{c} 1 \\ | \\ H \\ | \\ G \end{array} \right\}$$

given by the correspondences

$$E \longrightarrow \left\{ \begin{array}{c} \text{the elements of } G \\ \text{fixing } E \end{array} \right\}$$

$$\left\{ \begin{array}{c} \text{the fixed field} \\ \text{of } H \end{array} \right\} \longleftarrow H$$

which are inverse to each other. Under this correspondence:

(1) (inclusion reversing) If E_1, E_2 correspond to H_1, H_2, respectively, then $E_1 \subseteq E_2$ if and only if $H_2 \leq H_1$

(2) $[K : E] = |H|$ and $[E : F] = |G : H|$, the index of H in G:

$$\begin{array}{ll} K & \\ | & \} \quad |H| \\ E & \\ | & \} \quad |G : H| \\ F & \end{array}$$

(3) K/E is always Galois, with Galois group $\mathrm{Gal}(K/E) = H$:

$$\begin{array}{cc} K & \\ | & H \\ E & \end{array}$$

(4) E is Galois over F if and only if H is a normal subgroup in G. If this is the case, then the Galois group is isomorphic to the quotient group:

$$\mathrm{Gal}(E/F) \cong G/H.$$

More generally, even if H is not necessarily normal in G, the isomorphisms of E (into a fixed algebraic closure of F containing K) which fix F are in one to one correspondence with the cosets $\{\sigma H\}$ of H in G.

(5) If E_1, E_2 correspond to H_1, H_2, respectively, then the intersection $E_1 \cap E_2$ corresponds to the group $\langle H_1, H_2 \rangle$ generated by H_1 and H_2 and the composite field E_1E_2 corresponds to the intersection $H_1 \cap H_2$. Hence the lattice of subfields of K containing F and the lattice of subgroups of G are "dual" (the lattice diagram for one is the lattice diagram for the other turned upside down).

Proof: Given any subgroup H of G we obtain a unique fixed field $E = K_H$ by Corollary 12. This shows that the correspondence above is injective from right to left.

If K is the splitting field of the separable polynomial $f(x) \in F[x]$ then we may also view $f(x)$ as an element of $E[x]$ for any subfield E of K containing F. Then K is also the splitting field of $f(x)$ over E, so the extension K/E is Galois. By Corollary 10, E is the fixed field of $\mathrm{Aut}(K/E) \leq G$, showing that *every* subfield of K containing F arises as the fixed field for some subgroup of G. Hence the correspondence above is surjective from right to left, hence a bijection. The correspondences are inverse to each other since the automorphisms fixing E are precisely $\mathrm{Aut}(K/E)$ by Corollary 10.

We have already seen that the Galois correspondence is inclusion reversing in Proposition 4, which gives (1).

If $E = K_H$ is the fixed field of H, then Theorem 9 gives $[K : E] = |H|$ and $[K : F] = |G|$. Taking the quotient gives $[E : F] = |G : H|$, which proves (2).

Corollary 11 gives (3) immediately.

Suppose $E = K_H$ is the fixed field of the subgroup H. Every $\sigma \in G = \mathrm{Gal}(K/F)$ when restricted to E is an embedding $\sigma|_E$ of E with the subfield $\sigma(E)$ of K. Conversely, let $\tau : E \xrightarrow{\sim} \tau(E) \subseteq \overline{F}$ be any embedding of E (into a fixed algebraic closure $\overline{F}$ of F containing K) which fixes F. Then $\tau(E)$ is in fact contained in K: if $\alpha \in E$ has minimal polynomial $m_\alpha(x)$ over F then $\tau(\alpha)$ is another root of $m_\alpha(x)$ and K contains all these roots by Theorem 13. As above K is the splitting field of $f(x)$ over E and so also the splitting field of $\tau f(x)$ (which is the same as $f(x)$ since $f(x)$ has coefficients in F) over $\tau(E)$. Theorem 13.27 on extending isomorphisms then shows that we can extend τ to an isomorphism σ:

$$\begin{array}{cccc} \sigma: & K & \xrightarrow{\sim} & K \\ & | & & | \\ \tau: & E & \xrightarrow{\sim} & \tau(E). \end{array}$$

Since σ fixes F (because τ does), it follows that *every* embedding τ of E fixing F is the restriction to E of some automorphism σ of K fixing F, in other words, every embedding of E is of the form $\sigma|_E$ for some $\sigma \in G$.

Two automorphisms $\sigma, \sigma' \in G$ restrict to the *same* embedding of E if and only if $\sigma^{-1}\sigma'$ is the identity map on E. But then $\sigma^{-1}\sigma' \in H$ (i.e. $\sigma' \in \sigma H$) since by (3) the automorphisms of K which fix E are precisely the elements in H. Hence the distinct embeddings of E are in bijection with the cosets σH of H in G. In particular this gives

$$|\mathrm{Emb}(E/F)| = [G : H] = [E : F]$$

where $\mathrm{Emb}(E/F)$ denotes the set of embeddings of E (into a fixed algebraic closure of F) which fix F. Note that $\mathrm{Emb}(E/F)$ contains the automorphisms $\mathrm{Aut}(E/F)$.

The extension E/F will be Galois if and only if $|\mathrm{Aut}(E/F)| = [E : F]$. By the equality above, this will be the case if and only if each of the *embeddings* of E is actually an *automorphism* of E, i.e. if and only if $\sigma(E) = E$ for every $\sigma \in G$.

If $\sigma \in G$, then the subgroup of G fixing the field $\sigma(E)$ is the group $\sigma H \sigma^{-1}$, i.e.

$$\sigma(E) = K_{\sigma H \sigma^{-1}}.$$

To see this observe that if $\sigma\alpha \in \sigma(E)$ then

$$(\sigma h \sigma^{-1})(\sigma\alpha) = \sigma(h\alpha) = \sigma\alpha \qquad \text{for all } h \in H,$$

since h fixes $\alpha \in E$, which shows that $\sigma H \sigma^{-1}$ fixes $\sigma(E)$. The group fixing $\sigma(E)$ has order equal to the degree of K over $\sigma(E)$. But this is the same as the degree of K over E since the fields are isomorphic, hence the same as the order of H. Hence $\sigma H \sigma^{-1}$ is precisely the group fixing $\sigma(E)$ since we have shown containment and their orders are the same.

Because of the bijective nature of the Galois correspondence already proved we know that two subfields of K containing F are equal if and only if their fixing subgroups are equal in G. Hence $\sigma(E) = E$ for all $\sigma \in G$ if and only if $\sigma H \sigma^{-1} = H$ for all $\sigma \in G$, in other words E is Galois over F if and only if H is a normal subgroup of G.

We have already identified the embeddings of E over F as the set of cosets of H in G and when H is normal in G seen that the embeddings are automorphisms. It follows that in this case the *group* of cosets G/H is identified with the *group* of automorphisms of the Galois extension E/F by the definition of the group operation (composition of automorphisms). Hence $G/H \cong \mathrm{Gal}(E/F)$ when H is normal in G, which completes the proof of (4).

Suppose H_1 is the subgroup of elements of G fixing the subfield E_1 and H_2 is the subgroup of elements of G fixing the subfield E_2. Any element in $H_1 \cap H_2$ fixes both E_1 and E_2, hence fixes every element in the composite E_1E_2, since the elements in this field are algebraic combinations of the elements of E_1 and E_2. Conversely, if an automorphism σ fixes the composite E_1E_2 then in particular σ fixes E_1, i.e. $\sigma \in H_1$, and σ fixes E_2, i.e. $\sigma \in H_2$, hence $\sigma \in H_1 \cap H_2$. This proves that the composite E_1E_2 corresponds to the intersection $H_1 \cap H_2$. Similarly, the intersection $E_1 \cap E_2$ corresponds to the group $\langle H_1, H_2 \rangle$ generated by H_1 and H_2, completing the proof of the theorem.

Example: ($\mathbb{Q}(\sqrt{2}, \sqrt{3})$ **and** $\mathbb{Q}(\sqrt[3]{2}, \rho)$)

We have already seen examples of this theorem at the beginning of this section. We now see that the diagrams of subfields for the two fields $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ and $\mathbb{Q}(\sqrt[3]{2}, \rho)$ given above indicate *all* the subfields for these two fields.

Since every subgroup of the Klein 4-group is normal, all the subfields of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ are Galois extensions of $\mathbb{Q}$.

Similarly, since the only nontrivial normal subgroup of S_3 is the subgroup of order 3, we see that only the subfield $\mathbb{Q}(\rho)$ of $K = \mathbb{Q}(\sqrt[3]{2}, \rho)$ is Galois over $\mathbb{Q}$, with Galois group isomorphic to $S_3/\langle\sigma\rangle$, i.e. the cyclic group of order 2. For example, the nontrivial automorphism of $\mathbb{Q}(\rho)$ is induced by restricting any element (τ, for instance) in the nontrivial coset of $\langle\sigma\rangle$ to $\mathbb{Q}(\rho)$. This is clear from the explicit descriptions of these automorphisms given before—each of the elements $\tau, \tau\sigma, \tau\sigma^2$ in this coset map ρ to ρ^2. The restrictions of the elements of $\mathrm{Gal}(K/\mathbb{Q})$ to the (non-Galois) cubic subfields do not give automorphisms of these fields in general, rather giving isomorphisms of these fields with each other, in accordance with (4) of the theorem.

Example: ($\mathbb{Q}(\sqrt{2} + \sqrt{3})$)

Consider the field $\mathbb{Q}(\sqrt{2} + \sqrt{3})$. This is clearly a subfield of the Galois extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$. The other roots of the minimal polynomial for $\sqrt{2}+\sqrt{3}$ over $\mathbb{Q}$ are therefore the distinct conjugates of $\sqrt{2} + \sqrt{3}$ under the Galois group. The conjugates are

$$\pm\sqrt{2} \pm \sqrt{3}$$

which are easily seen to be distinct. The minimal polynomial is therefore

$$[x - (\sqrt{2} + \sqrt{3})][x - (\sqrt{2} - \sqrt{3})][x - (-\sqrt{2} + \sqrt{3})][x - (-\sqrt{2} - \sqrt{3})]$$

which is quickly computed to be the polynomial $x^4 - 10x^2 + 1$. It follows that this polynomial is irreducible and that

$$\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3}),$$

either by degree considerations or by noting that only the automorphism 1 of $\{1, \sigma, \tau, \sigma\tau\}$ fixes $\sqrt{2} + \sqrt{3}$ so the fixing group for this field is the same as for $\mathbb{Q}(\sqrt{2}, \sqrt{3})$.

Example: (Splitting field of $x^8 - 2$)

The splitting field of $x^8 - 2$ over $\mathbb{Q}$ is generated by $\theta = \sqrt[8]{2}$ (any fixed 8th root of 2, say the real one) and a primitive 8th root of unity $\zeta = \zeta_8$. Recall from Section 13.6 that

$$\mathbb{Q}(\zeta_8) = \mathbb{Q}(i, \sqrt{2}).$$

Since $\theta^4 = \sqrt{2}$ we see that the splitting field is generated by θ and i. The subfield $\mathbb{Q}(\theta)$ is of degree 8 over $\mathbb{Q}$ (since $x^8 - 2$ is irreducible, being Eisenstein), and all the elements of this field are real. Hence $i \notin \mathbb{Q}(\theta)$ and since i generates at most a quadratic extension of this field, the splitting field

$$\mathbb{Q}(\sqrt[8]{2}, \zeta_8) = \mathbb{Q}(\sqrt[8]{2}, i)$$

is of degree 16 over $\mathbb{Q}$.

The Galois group is determined by the action on the generators θ and i which gives the possibilities

$$\begin{cases} \theta \mapsto \zeta^a\theta & a = 0, 1, 2, \ldots, 7 \\ i \mapsto \pm i \end{cases}$$

Since we have already seen that the degree of the extension is 16 and there are only 16 possible such maps, it follows that in fact each of the maps above is an automorphism of $\mathbb{Q}(\sqrt[8]{2}, i)$ over $\mathbb{Q}$. Define the two automorphisms

$$\sigma : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto i \end{cases} \qquad \tau : \begin{cases} \theta \mapsto \theta \\ i \mapsto -i \end{cases}$$

(τ is the map induced by complex conjugation). Since

$$\begin{aligned} \zeta = \zeta_8 &= \frac{\sqrt{2}}{2} + i\frac{\sqrt{2}}{2} = \frac{1}{2}(1+i)\sqrt{2} \\ &= \frac{1}{2}(1+i)\theta^4 \end{aligned}$$

we can easily compute what happens to ζ from the explicit expressions for the powers of ζ in Fig.1. We find

$$\sigma : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta = \zeta^5 \end{cases} \qquad \tau : \begin{cases} \theta \mapsto \theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

Note that the reason we are interested in also keeping track of the action on the element ζ is that it will be needed in computing the composites of automorphisms, for example in computing

$$\begin{aligned} \sigma^2(\theta) = \sigma(\zeta\theta) = \sigma(\zeta)\sigma(\theta) &= (-\zeta)(\zeta\theta) = -\zeta^2\theta \\ &= -i\theta. \end{aligned}$$

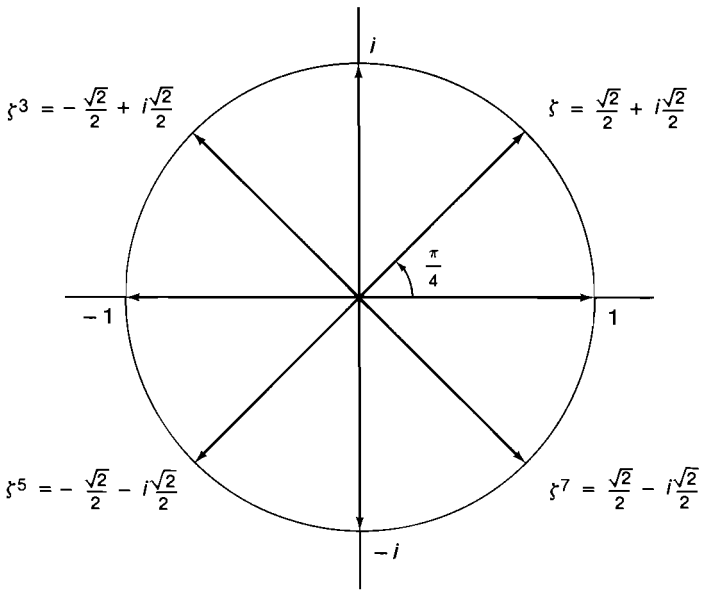


Fig. 1

We can similarly compute the following automorphisms:

$$\sigma : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto i \\ \zeta \mapsto \zeta^5 \end{cases} \qquad \tau\sigma : \begin{cases} \theta \mapsto \zeta^7\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

$$\sigma^2 : \begin{cases} \theta \mapsto \zeta^6\theta \\ i \mapsto i \\ \zeta \mapsto \zeta \end{cases} \qquad \tau\sigma^2 : \begin{cases} \theta \mapsto \zeta^2\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

$$\sigma^3 : \begin{cases} \theta \mapsto \zeta^7\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta \end{cases} \qquad \tau\sigma^3 : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

$$\sigma^4 : \begin{cases} \theta \mapsto -\theta \\ i \mapsto i \\ \zeta \mapsto \zeta \end{cases} \qquad \tau\sigma^4 : \begin{cases} \theta \mapsto -\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

$$\sigma^5 : \begin{cases} \theta \mapsto \zeta^5\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta \end{cases} \qquad \tau\sigma^5 : \begin{cases} \theta \mapsto \zeta^3\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

$$\sigma^6 : \begin{cases} \theta \mapsto \zeta^2\theta \\ i \mapsto i \\ \zeta \mapsto \zeta \end{cases} \qquad \tau\sigma^6 : \begin{cases} \theta \mapsto \zeta^6\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^7 \end{cases}$$

$$\sigma^7 : \begin{cases} \theta \mapsto \zeta^3\theta \\ i \mapsto i \\ \zeta \mapsto -\zeta \end{cases} \qquad \tau\sigma^7 : \begin{cases} \theta \mapsto \zeta^5\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3. \end{cases}$$

Since this exhausts the possibilities, these elements (together with 1 and τ) are the Galois group. We see in particular that σ and τ generate the Galois group. To determine the relations satisfied by these elements, we observe first that clearly $\tau^2 = 1$ and $(\sigma^4)^2 = 1$, so that

$$\sigma^8 = \tau^2 = 1.$$

Also, we compute

$$\sigma\tau : \begin{cases} \theta \mapsto \zeta\theta \\ i \mapsto -i \\ \zeta \mapsto \zeta^3 \end{cases}$$

so that

$$\sigma\tau = \tau\sigma^3.$$

It is not too difficult to show that these relations define the group completely, i.e.

$$\mathrm{Gal}(\mathbb{Q}(\sqrt[8]{2}, i)/\mathbb{Q}) = \langle\, \sigma, \tau \mid \sigma^8 = \tau^2 = 1, \sigma\tau = \tau\sigma^3 \,\rangle.$$

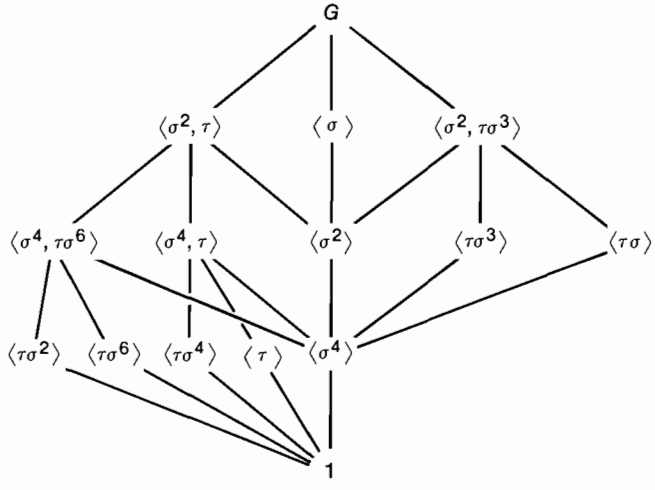
Such a group is called a *quasidihedral group* (recall that the dihedral group of order 16 would have the relation $\sigma\tau = \tau\sigma^7$ instead of $\sigma\tau = \tau\sigma^3$) and is a subgroup of S_8 since the Galois group is a subgroup of the permutations of the 8 roots of $x^8 - 2$.

This example again illustrates that one must take care in determining Galois groups from the actions on generators. We first computed the degree of the Galois extension above to determine the number of elements in the Galois group. Had we proceeded directly from the original generators $\theta = \sqrt[8]{2}$ and $\zeta = \zeta_8$ we might have (incorrectly) concluded that there were a total of 32 elements in the Galois group, since the first generator is mapped to any of 8 possible roots of $x^8 - 2$ and the second generator is mapped to any of 4 possible roots of its minimal polynomial $\Phi_4(x) = x^4 + 1$. The problem, as previously indicated, is that these choices are not independent. Here the reason is provided by the algebraic relation

$$\theta^4 = \sqrt{2} = \zeta + \zeta^7$$

which shows that one cannot specify the images of θ and ζ independently – their images must again satisfy this algebraic relation. This relation is perhaps sufficiently subtle to serve as a caution against rashly concluding maps are automorphisms. We note that in general it is necessary to provide justification that maps are automorphisms. This can be accomplished for example by using the extension theorems or by using degree considerations as we did here.

Determining the lattice of subgroups of this group G is a straightforward problem. The lattice is the following:



Determining the subfields corresponding to these subgroups (which by the Fundamental Theorem gives *all* the subfields of $\mathbb{Q}(\sqrt[8]{2}, i)$) is quite simple for a number of the

subgroups above using (2) of the Fundamental Theorem, which states that the degree of the extension over $\mathbb{Q}$ is equal to the *index* of the fixing subgroup. It then suffices to find a subfield of the right degree which is fixed by the subgroup in question. Remember also that if a subfield is fixed by the *generators* of a subgroup, then it is fixed by the subgroup. For example, from the explicit description for the automorphism σ we see that $\mathbb{Q}(i)$ is fixed by the group generated by σ. Since this is a subgroup of index 2 and $\mathbb{Q}(i)$ is of degree 2 over $\mathbb{Q}$, it must be the full fixed field. Most of the fixed fields for the subgroups above can be determined in as simple a manner.

For the subgroups of order 4 on the right (namely, generated by $\tau\sigma^3$ and by $\tau\sigma$), it is perhaps not so easy to see how to determine the corresponding fixed field. For the subgroup H generated by $\tau\sigma^3$ we may proceed as follows: the element $\theta^2 = \sqrt[4]{2}$ is clearly fixed by σ^4. By the diagram above, σ^4 is a normal subgroup of H of index 2, with representatives $1, \tau\sigma^3$ for the cosets. Consider the element

$$\alpha = (1 + \tau\sigma^3)\theta^2 = \theta^2 + \tau\sigma^3\theta^2.$$

Then α is fixed by σ^4 (we are in a commutative group H of order 4, so σ^4 commutes with 1 and $\tau\sigma^3$ and we already know θ^2 is fixed by σ^4). But (and this is the point), α is also fixed by $\tau\sigma^3$:

$$\begin{aligned}\tau\sigma^3\alpha &= \tau\sigma^3(1 + \tau\sigma^3)\theta^2 = [\tau\sigma^3 + (\tau\sigma^3)^2]\theta^2 \\ &= (\tau\sigma^3 + \sigma^4)\theta^2\end{aligned}$$

and the last expression is just α since $\sigma^4\theta^2 = \theta^2$. Hence α is an element of the fixed field for H. Explicitly

$$\alpha = \sqrt[4]{2} + i\sqrt[4]{2} = (1 + i)\sqrt[4]{2}.$$

A quick check shows that α is not fixed by the automorphism σ^2, so by the diagram of subgroups above, it follows that the fixing subgroup for the field $\mathbb{Q}(\alpha)$ is no larger than H, hence is precisely H, which gives us our fixed field. This also gives the fixed field for $\langle\tau\sigma\rangle$ by recalling that in general if E is the fixed field of H then the fixed field of $\tau H\tau^{-1}$ is the field $\tau(E)$. For $H = \langle\tau\sigma^3\rangle$, $\tau H\tau^{-1} = \langle\tau\sigma\rangle$, with fixed field given by $\tau(\alpha) = (1 - i)\sqrt[4]{2}$.

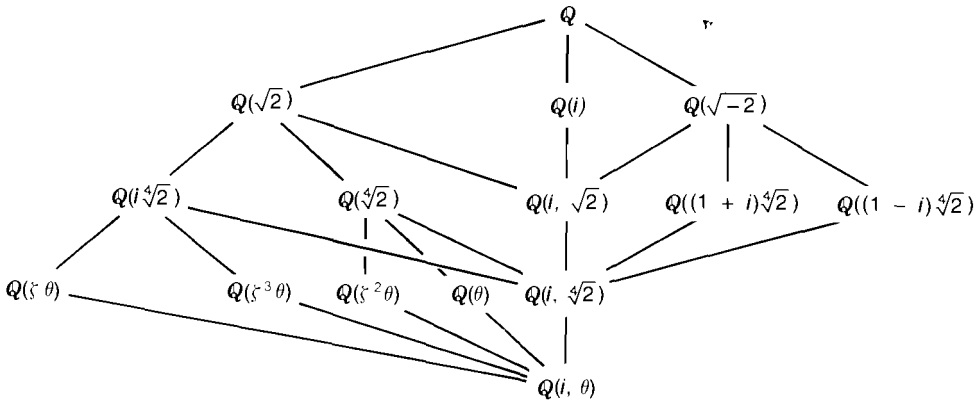
In general one tries to determine elements which are fixed by a given subgroup H of the Galois group (cf. the exercises, which indicate where the element above arose) and attempts to generate a sufficiently large field to give the full fixed field. In our case we were able to accomplish this with a single generator. We shall see later that every finite extension of $\mathbb{Q}$ is a simple extension, so there will be a single generator of this type, but in general it may be difficult to produce it directly.

The element α is a root of the polynomial

$$x^4 + 8$$

which must therefore be irreducible since we have already determined that a root of this polynomial generates an extension of degree 4 over $\mathbb{Q}$.

In a similar way it is possible to complete the diagram of subfields of $\mathbb{Q}(\sqrt[8]{2}, i)$, which we have inverted to emphasize its relation with the subgroup diagram above ($\theta = \sqrt[8]{2}$):



Note that the group $\langle\sigma^4\rangle$ is normal in G (in fact it is the center of G) with $G/\langle\sigma^4\rangle \cong D_8$, so the corresponding fixed field $\mathbb{Q}(i, \sqrt[4]{2})$ is Galois over $\mathbb{Q}$ with D_8 as Galois group. Being Galois it is a splitting field, evidently the splitting field for $x^4 - 2$. The lattice of subfields for this field is then immediate from the lattice above.

We end this example with the following amusing aspect of this Galois extension: It is an easy exercise to verify that

$$\langle\sigma^2, \tau\rangle \cong D_8 \quad \langle\sigma\rangle \cong \mathbb{Z}/8\mathbb{Z} \quad \langle\sigma^2, \tau\sigma^3\rangle \cong Q_8$$

where D_8 is the dihedral group of order 8 and Q_8 is the quaternion group of order 8. It follows that the field $\mathbb{Q}(\sqrt[8]{2}, i)$ is Galois of degree 8 over its three quadratic subfields

$$\mathbb{Q}(\sqrt{2}) \qquad \mathbb{Q}(i) \qquad \mathbb{Q}(\sqrt{-2})$$

with dihedral, cyclic and quaternion Galois groups, respectively, so that three of the 5 possible groups of order 8 (and both non-abelian ones) appear as Galois groups in this extension.

We shall consider additional examples and applications in the following sections.

EXERCISES

1. Determine the minimal polynomial over $\mathbb{Q}$ for the element $\sqrt{2} + \sqrt{5}$.
2. Determine the minimal polynomial over $\mathbb{Q}$ for the element $1 + \sqrt[3]{2} + \sqrt[3]{4}$.
3. Determine the Galois group of $(x^2 - 2)(x^2 - 3)(x^2 - 5)$. Determine *all* the subfields of the splitting field of this polynomial.
4. Let p be a prime. Determine the elements of the Galois group of $x^p - 2$.
5. Prove that the Galois group of $x^p - 2$ for p a prime is isomorphic to the group of matrices $\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ where $a, b \in \mathbb{F}_p$, $a \neq 0$.
6. Let $K = \mathbb{Q}(\sqrt[8]{2}, i)$ and let $F_1 = \mathbb{Q}(i)$, $F_2 = \mathbb{Q}(\sqrt{2})$, $F_3 = \mathbb{Q}(\sqrt{-2})$. Prove that $\mathrm{Gal}(K/F_1) \cong Z_8$, $\mathrm{Gal}(K/F_2) \cong D_8$, $\mathrm{Gal}(K/F_3) \cong Q_8$.
7. Determine all the subfields of the splitting field of $x^8 - 2$ which are Galois over $\mathbb{Q}$.

8. Suppose K is a Galois extension of F of degree p^n for some $n \geq 1$. Show there are Galois extensions of F contained in K of degrees p and p^{n-1}.
9. Give an example of fields F_1, F_2, F_3 with $\mathbb{Q} \subset F_1 \subset F_2 \subset F_3$, $[F_3 : \mathbb{Q}] = 8$ and each field is Galois over all its subfields with the exception that F_2 is not Galois over $\mathbb{Q}$.
10. Determine the Galois group of the splitting field over $\mathbb{Q}$ of $x^8 - 3$.
11. Suppose $f(x) \in \mathbb{Z}[x]$ is an irreducible quartic whose splitting field has Galois group S_4 over $\mathbb{Q}$ (there are many such quartics, cf. Section 6). Let θ be a root of $f(x)$ and set $K = \mathbb{Q}(\theta)$. Prove that K is an extension of $\mathbb{Q}$ of degree 4 which has no proper subfields. Are there any Galois extensions of $\mathbb{Q}$ of degree 4 with no proper subfields?
12. Determine the Galois group of the splitting field over $\mathbb{Q}$ of $x^4 - 14x^2 + 9$.
13. Prove that if the Galois group of the splitting field of a cubic over $\mathbb{Q}$ is the cyclic group of order 3 then all the roots of the cubic are real.
14. Show that $\mathbb{Q}(\sqrt{2+\sqrt{2}})$ is a cyclic quartic field, i.e. is a Galois extension of degree 4 with cyclic Galois group.
15. (*Biquadratic Extensions*) Let F be a field of characteristic $\neq 2$.
    (a) If $K = F(\sqrt{D_1}, \sqrt{D_2})$ where $D_1, D_2 \in F$ have the property that none of D_1, D_2 or D_1D_2 is a square in F, prove that K/F is a Galois extension with $\mathrm{Gal}(K/F)$ isomorphic to the Klein 4-group.
    (b) Conversely, suppose K/F is a Galois extension with $\mathrm{Gal}(K/F)$ isomorphic to the Klein 4-group. Prove that $K = F(\sqrt{D_1}, \sqrt{D_1})$ where $D_1, D_2 \in F$ have the property that none of D_1, D_2 or D_1D_2 is a square in F.
16. (a) Prove that $x^4 - 2x^2 - 2$ is irreducible over $\mathbb{Q}$.
    (b) Show the roots of this quartic are

$$\alpha_1 = \sqrt{1+\sqrt{3}} \qquad \alpha_3 = -\sqrt{1+\sqrt{3}}$$
$$\alpha_2 = \sqrt{1-\sqrt{3}} \qquad \alpha_4 = -\sqrt{1-\sqrt{3}}.$$

    (c) Let $K_1 = \mathbb{Q}(\alpha_1)$ and $K_2 = \mathbb{Q}(\alpha_2)$. Show that $K_1 \neq K_2$ and that $K_1 \cap K_2 = \mathbb{Q}(\sqrt{3}) = F$.
    (d) Prove that K_1, K_2 and K_1K_2 are Galois over F with $\mathrm{Gal}(K_1K_2/F)$ the Klein 4-group. Write out the elements of $\mathrm{Gal}(K_1K_2/F)$ explicitly. Determine all the subgroups of the Galois group and give their corresponding fixed subfields of K_1K_2 containing F.
    (e) Prove that the splitting field of $x^4 - 2x^2 - 2$ over $\mathbb{Q}$ is of degree 8 with dihedral Galois group.

The following two exercises indicate one method for constructing elements in subfields of a given field and are quite useful in many computations.

17. Let K/F be any finite extension and let $\alpha \in K$. Let L be a Galois extension of F containing K and let $H \leq \mathrm{Gal}(L/F)$ be the subgroup corresponding to K. Define the *norm* of α from K to F to be

$$\mathrm{N}_{K/F}(\alpha) = \prod_{\sigma} \sigma(\alpha),$$

where the product is taken over all the embeddings of K into an algebraic closure of F (so over a set of coset representatives for H in $\mathrm{Gal}(L/F)$ by the Fundamental Theorem of Galois Theory). This is a product of Galois conjugates of α. In particular, if K/F is Galois this is $\prod_{\sigma \in \mathrm{Gal}(K/F)} \sigma(\alpha)$.

(a) Prove that $N_{K/F}(\alpha) \in F$.

(b) Prove that $N_{K/F}(\alpha\beta) = N_{K/F}(\alpha)N_{K/F}(\beta)$, so that the norm is a multiplicative map from K to F.

(c) Let $K = F(\sqrt{D})$ be a quadratic extension of F. Show that $N_{K/F}(a + b\sqrt{D}) = a^2 - Db^2$.

(d) Let $m_\alpha(x) = x^d + a_{d-1}x^{d-1} + \cdots + a_1x + a_0 \in F[x]$ be the minimal polynomial for $\alpha \in K$ over F. Let $n = [K : F]$. Prove that d divides n, that there are d distinct Galois conjugates of α which are all repeated n/d times in the product above and conclude that $N_{K/F}(\alpha) = (-1)^n a_0^{n/d}$.

18. With notation as in the previous problem, define the *trace* of α from K to F to be

$$\mathrm{Tr}_{K/F}(\alpha) = \sum_\sigma \sigma(\alpha),$$

a sum of Galois conjugates of α.

(a) Prove that $\mathrm{Tr}_{K/F}(\alpha) \in F$.

(b) Prove that $\mathrm{Tr}_{K/F}(\alpha+\beta) = \mathrm{Tr}_{K/F}(\alpha) + \mathrm{Tr}_{K/F}(\beta)$, so that the trace is an additive map from K to F.

(c) Let $K = F(\sqrt{D})$ be a quadratic extension of F. Show that $\mathrm{Tr}_{K/F}(a + b\sqrt{D}) = 2a$.

(d) Let $m_\alpha(x)$ be as in the previous problem. Prove that $\mathrm{Tr}_{K/F}(\alpha) = -\dfrac{n}{d}a_{d-1}$.

19. With notation as in the previous problems show that $N_{K/F}(a\alpha) = a^n N_{K/F}(\alpha)$ and $\mathrm{Tr}_{K/F}(a\alpha) = a\mathrm{Tr}_{K/F}(\alpha)$ for all a in the base field F. In particular show that $N_{K/F}(a) = a^n$ and $\mathrm{Tr}_{K/F}(a) = na$ for all $a \in F$.

20. With notation as in the previous problems show more generally that $\prod_\sigma(x - \sigma(\alpha)) = (m_\alpha(x))^{n/d}$.

21. Use the linear independence of characters to show that for any Galois extension K of F there is an element $\alpha \in K$ with $\mathrm{Tr}_{K/F}(\alpha) \neq 0$.

22. Suppose K/F is a Galois extension and let σ be an element of the Galois group.

(a) Suppose $\alpha \in K$ is of the form $\alpha = \dfrac{\beta}{\sigma\beta}$ for some nonzero $\beta \in K$. Prove that $N_{K/F}(\alpha) = 1$.

(b) Suppose $\alpha \in K$ is of the form $\alpha = \beta - \sigma\beta$ for some $\beta \in K$. Prove that $\mathrm{Tr}_{K/F}(\alpha) = 0$.

23. (*Hilbert's Theorem 90*) Let K be a Galois extension of F with cyclic Galois group of order n. Suppose $\alpha \in K$ has $N_{K/F}(\alpha) = 1$. Prove that α is of the form $\alpha = \dfrac{\beta}{\sigma\beta}$ for some nonzero $\beta \in K$. [By the linear independence of characters show there exists some $\theta \in K$ such that

$$\beta = \theta + \alpha\sigma(\theta) + (\alpha\,\sigma\alpha)\sigma^2(\theta) + \cdots + (\alpha\,\sigma\alpha\ldots\sigma^{n-2}\alpha)\sigma^{n-1}(\theta)$$

is nonzero. Compute $\dfrac{\beta}{\sigma\beta}$ using the fact that α has norm 1 to F.]

24. Prove that the rational solutions $a, b \in \mathbb{Q}$ of Pythagoras' equation $a^2 + b^2 = 1$ are of the form $a = \dfrac{s^2 - t^2}{s^2 + t^2}$ and $b = \dfrac{2st}{s^2 + t^2}$ for some $s, t \in \mathbb{Q}$ and hence show that any right triangle with integer sides has sides of lengths $(m^2 - n^2, 2mn, m^2 + n^2)$ for some integers m, n. [Note that $a^2 + b^2 = 1$ is equivalent to $N_{\mathbb{Q}(i)/\mathbb{Q}}(a + ib) = 1$, then use Hilbert's Theorem 90 above with $\beta = s + it$.]

25. Generalize the previous problem to determine all the rational solutions of the equation $a^2 + Db^2 = 1$ for $D \in \mathbb{Z}$, $D > 0$, D not a perfect square in $\mathbb{Z}$.

26. Let K be a Galois extension of F with cyclic Galois group of order n. Suppose $\alpha \in K$ has $\mathrm{Tr}_{K/F}(\alpha) = 0$. Prove that α is of the form $\alpha = \beta - \sigma\beta$ for some $\beta \in K$. [Let $\theta \in K$ be an element with $\mathrm{Tr}_{K/F}(\theta) \neq 0$ by a previous exercise, let

$$\beta = \frac{1}{\mathrm{Tr}_{K/F}(\theta)}[\alpha\sigma(\theta) + (\alpha + \sigma\alpha)\sigma^2(\theta) + \cdots + (\alpha + \sigma\alpha + \cdots + \sigma^{n-2}\alpha)\sigma^{n-1}(\theta)]$$

and compute $\beta - \sigma\beta$.]

27. Let $\alpha = \sqrt{(2+\sqrt{2})(3+\sqrt{3})}$ (positive real square roots for concreteness) and consider the extension $E = \mathbb{Q}(\alpha)$.

(a) Show that $a = (2+\sqrt{2})(3+\sqrt{3})$ is not a square in $F = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. [If $a = c^2, c \in F$, then $a\varphi(a) = (2+\sqrt{2})^2(6) = (c\varphi c)^2$ for the automorphism $\varphi \in \mathrm{Gal}(F/\mathbb{Q})$ fixing $\mathbb{Q}(\sqrt{2})$. Since $c\varphi c = \mathrm{N}_{F/\mathbb{Q}(\sqrt{2})}(c) \in \mathbb{Q}(\sqrt{2})$ conclude that this implies $\sqrt{6} \in \mathbb{Q}(\sqrt{2})$, a contradiction.]

(b) Conclude from (a) that $[E : \mathbb{Q}] = 8$. Prove that the roots of the minimal polynomial over $\mathbb{Q}$ for α are the 8 elements $\pm\sqrt{(2\pm\sqrt{2})(3\pm\sqrt{3})}$.

(c) Let $\beta = \sqrt{(2-\sqrt{2})(3+\sqrt{3})}$. Show that $\alpha\beta = \sqrt{2}(3+\sqrt{3}) \in F$ so that $\beta \in E$. Show similarly that the other roots are also elements of E so that E is a Galois extension of $\mathbb{Q}$. Show that the elements of the Galois group are precisely the maps determined by mapping α to one of the eight elements in (b).

(d) Let $\sigma \in \mathrm{Gal}(E/\mathbb{Q})$ be the automorphism which maps α to β. Show that since $\sigma(\alpha^2) = \beta^2$ that $\sigma(\sqrt{2}) = -\sqrt{2}$ and $\sigma(\sqrt{3}) = \sqrt{3}$. From $\alpha\beta = \sqrt{2}(3+\sqrt{3})$ conclude that $\sigma(\alpha\beta) = -\alpha\beta$ and hence $\sigma(\beta) = -\alpha$. Show that σ is an element of order 4 in $\mathrm{Gal}(E/\mathbb{Q})$.

(e) Show similarly that the map τ defined by $\tau(\alpha) = \sqrt{(2+\sqrt{2})(3-\sqrt{3})}$ is an element of order 4 in $\mathrm{Gal}(E/\mathbb{Q})$. Prove that σ and τ generate the Galois group, $\sigma^4 = \tau^4 = 1$, $\sigma^2 = \tau^2$ and that $\sigma\tau = \tau\sigma^3$.

(f) Conclude that $\mathrm{Gal}(E/\mathbb{Q}) \cong Q_8$, the quaternion group of order 8.

28. Let $f(x) \in F[x]$ be an irreducible polynomial of degree n over the field F, let L be the splitting field of $f(x)$ over F and let α be a root of $f(x)$ in L. If K is any Galois extension of F contained in L, show that the polynomial $f(x)$ splits into a product of m irreducible polynomials each of degree d over K, where $m = [F(\alpha) \cap K : F]$ and $d = [K(\alpha) : K]$ (cf. also the generalization in Exercise 4 of Section 4). [If H is the subgroup of the Galois group of L over F corresponding to K then the factors of $f(x)$ over K correspond to the orbits of H on the roots of $f(x)$. Then use Exercise 9 of Section 4.1.]

29. Let F be a field and let $F(x)$ be the field of rational functions in the variable x. Define the maps σ and τ of $F(x)$ to itself by $\sigma f(x) = f(\frac{1}{1-x})$ and $\tau f(x) = f(\frac{1}{x})$ for $f(x) \in F(x)$.

(a) Prove that σ and τ are automorphisms of $F(x)$ (cf. Exercise 8 of Section 1) and that the group $G = \langle \sigma, \tau \rangle$ they generate is isomorphic to S_3.

(b) Prove that the element $t = \dfrac{(x^2-x+1)^3}{x^2(x-1)^2}$ is fixed by all the elements of G.

(c) Prove that $F(t)$ is precisely the fixed field of G in $F(x)$ [compute the degree of the extension].

30. Prove that the fixed field of the subgroup of automorphisms generated by τ in the previous problem is $F(x + \frac{1}{x})$. Prove that the fixed field of the subgroup generated by the automorphism $\tau\sigma^2$ (which maps x to $1-x$) is $F(x(1-x))$. Determine the fixed field of the subgroup generated by $\tau\sigma$ and the fixed field of the subgroup generated by σ.

31. Let K be a finite extension of F of degree n. Let α be an element of K.
   (a) Prove that α acting by left multiplication on K is an F-linear transformation T_α of K.
   (b) Prove that the minimal polynomial for α over F is the same as the minimal polynomial for the linear transformation T_α.
   (c) Prove that the trace $\mathrm{Tr}_{K/F}(\alpha)$ is the trace of the $n \times n$ matrix defined by T_α (which justifies these two uses of the same word "trace"). Prove that the norm $\mathrm{N}_{K/F}(\alpha)$ is the determinant of T_α.

14.3 FINITE FIELDS

A finite field $\mathbb{F}$ has characteristic p for some prime p so is a finite dimensional vector space over $\mathbb{F}_p$. If the dimension is n, i.e. $[\mathbb{F} : \mathbb{F}_p] = n$, then $\mathbb{F}$ has precisely p^n elements. We have already seen (following Proposition 13.37) that $\mathbb{F}$ is then isomorphic to the splitting field of the polynomial $x^{p^n} - x$, hence is unique up to isomorphism. We denote the finite field of order p^n by $\mathbb{F}_{p^n}$.

The field $\mathbb{F}_{p^n}$ is Galois over $\mathbb{F}_p$, with cyclic Galois group of order n generated by the Frobenius automorphism

$$\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p) = \langle \sigma_p \rangle \cong \mathbb{Z}/n\mathbb{Z}$$

where

$$\begin{aligned} \sigma_p : \mathbb{F}_{p^n} &\to \mathbb{F}_{p^n} \\ \alpha &\mapsto \alpha^p \end{aligned}$$

(Example 7 following Corollary 6). By the Fundamental Theorem, every subfield of $\mathbb{F}_{p^n}$ corresponds to a subgroup of $\mathbb{Z}/n\mathbb{Z}$. Hence for every divisor d of n there is precisely one subfield of $\mathbb{F}_{p^n}$ of degree d over $\mathbb{F}_p$, namely the fixed field of the subgroup generated by ${\sigma_p}^d$ of order n/d, and there are no other subfields. This field is isomorphic to $\mathbb{F}_{p^d}$, the unique finite field of order p^d.

Since the Galois group is abelian, every subgroup is normal, so each of the subfields $\mathbb{F}_{p^d}$ (d a divisor of n) is Galois over $\mathbb{F}_p$ (which is also clear from the fact that these are themselves splitting fields). Further, the Galois group $\mathrm{Gal}(\mathbb{F}_{p^d}/\mathbb{F}_p)$ is generated by the image of σ_p in the quotient group $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)/\langle {\sigma_p}^d \rangle$. If we denote this element again by σ_p, we recover the Frobenius automorphism for the extension $\mathbb{F}_{p^d}/\mathbb{F}_p$. (Note, however, that σ_p has order n in $\mathrm{Gal}(\mathbb{F}_{p^n}/\mathbb{F}_p)$ and order d in $\mathrm{Gal}(\mathbb{F}_{p^d}/\mathbb{F}_p)$.)

We summarize this as:

Proposition 15. Any finite field is isomorphic to $\mathbb{F}_{p^n}$ for some prime p and some integer $n \geq 1$. The field $\mathbb{F}_{p^n}$ is the splitting field of the polynomial $x^{p^n} - x$, with cyclic Galois group of order n generated by the Frobenius automorphism σ_p. The subfields of $\mathbb{F}_{p^n}$ are all Galois over $\mathbb{F}_p$ and are in one to one correspondence with the divisors d of n. They are the fields $\mathbb{F}_{p^d}$, the fixed fields of ${\sigma_p}^d$.

The corresponding statements for the finite extensions of any finite field are easy consequences of Proposition 15 and are outlined in the exercises.

As an elementary application we have the following result on the polynomial $x^4 + 1$ in $\mathbb{Z}[x]$:

Corollary 16. The irreducible polynomial $x^4 + 1 \in \mathbb{Z}[x]$ is reducible modulo every prime p.

Proof: Consider the polynomial $x^4 + 1$ over $\mathbb{F}_p[x]$ for the prime p. If $p = 2$ we have $x^4 + 1 = (x + 1)^4$ and the polynomial is reducible. Assume now that p is odd. Then $p^2 - 1$ is divisible by 8 since p is congruent mod 8 to 1,3,5 or 7 and all of these square to 1 mod 8. Hence $x^{p^2-1} - 1$ is divisible by $x^8 - 1$. Then we have the divisibilities

$$x^4 + 1 \mid x^8 - 1 \mid x^{p^2-1} - 1 \mid x^{p^2} - x$$

which shows that all the roots of $x^4 + 1$ are roots of $x^{p^2} - x$. (Equivalently, these roots are fixed by the square of the Frobenius automorphism ${\sigma_p}^2$.) Since the roots of $x^{p^2} - x$ are the field $\mathbb{F}_{p^2}$, it follows that the extension generated by any root of $x^4 + 1$ is at most of degree 2 over $\mathbb{F}_p$, which means that $x^4 + 1$ cannot be irreducible over $\mathbb{F}_p$.

The multiplicative group ${\mathbb{F}_{p^n}}^\times$ is obviously a finite subgroup of the multiplicative group of a field. By Proposition 9.18, this is a *cyclic* group. If θ is any generator, then clearly $\mathbb{F}_{p^n} = \mathbb{F}_p(\theta)$. This proves:

Proposition 17. The finite field $\mathbb{F}_{p^n}$ is simple. In particular, there exists an irreducible polynomial of degree n over $\mathbb{F}_p$ for every $n \geq 1$.

We have described the finite fields $\mathbb{F}_{p^n}$ above as the splitting fields of the polynomials $x^{p^n} - x$. By the previous proposition, this field can also be described as a quotient of $\mathbb{F}_p[x]$, namely by the minimal polynomial for θ. Since θ is necessarily a root of $x^{p^n} - x$, we see that the minimal polynomial for θ is a divisor of $x^{p^n} - x$ of degree n.

Conversely, let $p(x)$ be any irreducible polynomial of degree d, say, dividing $x^{p^n} - x$. If α is a root of $p(x)$, then the extension $\mathbb{F}_p(\alpha)$ is a subfield of $\mathbb{F}_{p^n}$ of degree d. Hence d is a divisor of n and the extension is Galois by Proposition 15 (in fact, the extension $\mathbb{F}_{p^d}$) so in particular all the roots of $p(x)$ are contained in $\mathbb{F}_p(\alpha)$.

The elements of $\mathbb{F}_{p^n}$ are precisely the roots of $x^{p^n} - x$. If we group together the factors $x - \alpha$ of this polynomial according to the degree d of their minimal polynomials over $\mathbb{F}_p$, we obtain

Proposition 18. The polynomial $x^{p^n} - x$ is precisely the product of all the distinct irreducible polynomials in $\mathbb{F}_p[x]$ of degree d where d is a divisor of n.

This proposition can be used to produce irreducible polynomials over $\mathbb{F}_p$ recursively. For example, the irreducible quadratics over $\mathbb{F}_2$ are the divisors of

$$\frac{x^4 - x}{x(x-1)}$$

which gives the single polynomial $x^2 + x + 1$. Similarly, the irreducible cubics over this field are the divisors of

$$\frac{x^8 - x}{x(x-1)} = x^6 + x^5 + x^4 + x^3 + x^2 + x + 1$$

which factors into the two cubics $x^3 + x + 1$ and $x^3 + x^2 + 1$. The irreducible quartics are given by dividing $x^{16} - x$ by $x(x-1)$ and the irreducible quadratic $x^2 + x + 1$ above and then factoring into irreducible quartics:

$$\frac{x^{16} - x}{x(x-1)(x^2 + x + 1)} = (x^4 + x^3 + x^2 + x + 1)(x^4 + x^3 + 1)(x^4 + x + 1).$$

This gives a method for determining the product of all the irreducible polynomials over $\mathbb{F}_p$ of a given degree. There exist efficient algorithms for factorization of polynomials mod p which will give the individual irreducible polynomials (cf. the exercises) in practice. The importance of having irreducible polynomials at hand is that they give a representation of the finite fields $\mathbb{F}_{p^n}$ (as quotients $\mathbb{F}_p[x]/(f(x))$ for $f(x)$ irreducible of degree n) conducive to explicit computations.

Note also that since the finite field $\mathbb{F}_{p^n}$ is unique up to isomorphism, the quotients of $\mathbb{F}_p[x]$ by any of the irreducible polynomials of degree n are all isomorphic. If $f_1(x)$ and $f_2(x)$ are irreducible of degree n, then $f_2(x)$ splits completely in the field $\mathbb{F}_{p^n} \cong \mathbb{F}_p[x]/(f_1(x))$. If we denote a root of $f_2(x)$ by $\alpha(x)$ (to emphasize that it is a polynomial of degree $< n$ in x in $\mathbb{F}_p[x]/(f_1(x))$), then the isomorphism is given by

$$\begin{gathered} \mathbb{F}_p[x]/(f_2(x)) \cong \mathbb{F}_p[x]/(f_1(x)) \\ x \mapsto \alpha(x) \end{gathered}$$

(we have mapped a root of $f_2(x)$ in the first field to a root of $f_2(x)$ in the second field). For example, if $f_1(x) = x^4 + x^3 + 1$, $f_2(x) = x^4 + x + 1$ are two of the irreducible quartics over $\mathbb{F}_2$ determined above, then a simple computation verifies that

$$\alpha(x) = x^3 + x^2$$

is a root of $f_2(x)$ in $\mathbb{F}_{16} = \mathbb{F}_2[x]/(x^4 + x^3 + 1)$. Then we have

$$\begin{gathered} \mathbb{F}_2[x]/(x^4 + x + 1) \cong \mathbb{F}_2[x]/(x^4 + x^3 + 1) \quad (\cong \mathbb{F}_{16}) \\ x \mapsto x^3 + x^2. \end{gathered}$$

If we assume a result from elementary number theory we can give a formula for the number of irreducible polynomials of degree n. Define the *Möbius* μ-function by

$$\mu(n) = \begin{cases} 1 & \text{for } n = 1 \\ 0 & \text{if } n \text{ has a square factor} \\ (-1)^r & \text{if } n \text{ has } r \text{ distinct prime factors.} \end{cases}$$

If now $f(n)$ is a function defined for all nonnegative integers n and $F(n)$ is defined by

$$F(n) = \sum_{d|n} f(d) \qquad n = 1, 2, \ldots$$

then the *Möbius inversion formula* states that one can recover the function $f(n)$ from $F(n)$:

$$f(n) = \sum_{d|n} \mu(d) F(\frac{n}{d}) \qquad n = 1, 2, \ldots.$$

This is an elementary result from number theory which we take for granted. Define

$$\psi(n) = \text{the number of irreducible polynomials of degree } n \text{ in } \mathbb{F}_p[x].$$

Counting degrees in Proposition 18 we have

$$p^n = \sum_{d|n} d\psi(d).$$

Applying the Möbius inversion formula (for $f(n) = n\psi(n)$) we obtain:

$$n\psi(n) = \sum_{d|n} \mu(d) p^{n/d}$$

which gives us a formula for the number of irreducible polynomials of degree n over $\mathbb{F}_p$:

$$\psi(n) = \frac{1}{n} \sum_{d|n} \mu(d) p^{n/d}.$$

For example, in the case $p = 2$, $n = 4$ we have

$$\psi(4) = \frac{1}{4}[\mu(1)2^4 + \mu(2)2^2 + \mu(4)2^1] = \frac{1}{4}(16 - 4 + 0) = 3$$

as we determined directly above.

We have seen above that

$$\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n} \text{ if and only if } m \text{ divides } n.$$

In particular, given any two finite fields $\mathbb{F}_{p^{n_1}}$ and $\mathbb{F}_{p^{n_2}}$ there is a third finite field containing (an isomorphic copy of) them, namely $\mathbb{F}_{p^{n_1 n_2}}$. This gives us a partial ordering on these fields and allows us to think of their union. Since these give *all* the finite extensions of $\mathbb{F}_p$, we see that the union of $\mathbb{F}_{p^n}$ for all n is an algebraic closure of $\mathbb{F}_p$, unique up to isomorphism:

$$\overline{\mathbb{F}_p} = \bigcup_{n \geq 1} \mathbb{F}_{p^n}.$$

This provides a simple description of the algebraic closure of $\mathbb{F}_p$.

EXERCISES

1. Factor $x^8 - x$ into irreducibles in $\mathbb{Z}[x]$ and in $\mathbb{F}_2[x]$.
2. Write out the multiplication table for $\mathbb{F}_4$ and $\mathbb{F}_8$.
3. Prove that an algebraically closed field must be infinite.

4. Construct the finite field of 16 elements and find a generator for the multiplicative group. How many generators are there?
5. Exhibit an explicit isomorphism between the splitting fields of $x^3 - x + 1$ and $x^3 - x - 1$ over $\mathbb{F}_3$.
6. Suppose $K = \mathbb{Q}(\theta) = \mathbb{Q}(\sqrt{D_1}, \sqrt{D_2})$ with $D_1, D_2 \in \mathbb{Z}$, is a biquadratic extension and that $\theta = a + b\sqrt{D_1} + c\sqrt{D_2} + d\sqrt{D_1 D_2}$ where $a, b, c, d \in \mathbb{Z}$ are integers. Prove that the minimal polynomial $m_\theta(x)$ for θ over $\mathbb{Q}$ is irreducible of degree 4 over $\mathbb{Q}$ but is reducible modulo every prime p. In particular show that the polynomial $x^4 - 10x^2 + 1$ is irreducible in $\mathbb{Z}[x]$ but is reducible modulo every prime. [Use the fact that there are no biquadratic extensions over finite fields.]
7. Prove that one of 2,3 or 6 is a square in $\mathbb{F}_p$ for every prime p. Conclude that the polynomial
$$x^6 - 11x^4 + 36x^2 - 36 = (x^2 - 2)(x^2 - 3)(x^2 - 6)$$
has a root modulo p for every prime p but has no root in $\mathbb{Z}$.
8. Determine the splitting field of the polynomial $x^p - x - a$ over $\mathbb{F}_p$ where $a \neq 0$, $a \in \mathbb{F}_p$. Show explicitly that the Galois group is cyclic. [Show $\alpha \mapsto \alpha + 1$ is an automorphism.] Such an extension is called an *Artin–Schreier extension* (cf. Exercise 9 of Section 7).
9. Let $q = p^m$ be a power of the prime p and let $\mathbb{F}_q = \mathbb{F}_{p^m}$ be the finite field with q elements. Let $\sigma_q = \sigma_p^m$ be the mth power of the Frobenius automorphism σ_p, called the q–Frobenius automorphism.
   (a) Prove that σ_q fixes $\mathbb{F}_q$.
   (b) Prove that every finite extension of $\mathbb{F}_q$ of degree n is the splitting field of $x^{q^n} - x$ over $\mathbb{F}_q$, hence is unique.
   (c) Prove that every finite extension of $\mathbb{F}_q$ of degree n is cyclic with σ_q as generator.
   (d) Prove that the subfields of the unique extension of $\mathbb{F}_q$ of degree n are in bijective correspondence with the divisors d of n.
10. Prove that n divides $\varphi(p^n - 1)$. [Observe that $\varphi(p^n - 1)$ is the order of the group of automorphisms of a cyclic group of order $p^n - 1$.]
11. Prove that $x^{p^n} - x + 1$ is irreducible over $\mathbb{F}_p$ only when $n = 1$ or $n = p = 2$. [Note that if α is a root, then so is $\alpha + a$ for any $a \in \mathbb{F}_{p^n}$. Show that this implies $\mathbb{F}_p(\alpha)$ contains $\mathbb{F}_{p^n}$ and that $[\mathbb{F}_p(\alpha) : \mathbb{F}_{p^n}] = p$.]

(*Berlekamp's Factorization Algorithm*) The following exercises outline the Berlekamp factorization algorithm for factoring polynomials in $\mathbb{F}_p[x]$. The efficiency of this algorithm is based on the efficiency of computing greatest common divisors in $\mathbb{F}_p[x]$ by the Euclidean Algorithm and on the efficiency of row-reduction matrix algorithms for solving systems of linear equations.

Let $f(x) \in \mathbb{F}_p[x]$ be a monic polynomial of degree n and let $f(x) = p_1(x)p_2(x)\dots p_k(x)$ where $p_1(x), p_2(x), \dots, p_k(x)$ are powers of distinct monic irreducibles in $\mathbb{F}_p[x]$.

12. Show that in order to write $f(x)$ as a product of irreducible polynomials in $\mathbb{F}_p[x]$ it suffices to determine the factors $p_1(x), \dots, p_k(x)$. [If $p(x) = q(x)^N \in \mathbb{F}_p[x]$ with $q(x)$ monic and irreducible, show that $q(x)$ can be determined from $p(x)$ by checking for p^{th} powers and by computing greatest common divisors with derivatives.]
13. Let $g(x) \in \mathbb{F}_p[x]$ be any polynomial of degree $< n$. Denote by $R(h(x))$ the remainder of $h(x)$ after division by $f(x)$. Prove the following are equivalent:

(a) $R(g(x^p)) = g(x)$.
(b) $f(x)$ divides $[g(x) - 0][g(x) - 1] \dots [g(x) - (p-1)]$. [Use the fact that $g(x^p) = g(x)^p$ together with the factorization of $x^p - x$ in $\mathbb{F}_p[x]$.]
(c) $p_i(x)$ divides the product in (b) for $i = 1, 2, \dots, k$.
(d) For each i, $i = 1, 2, \dots, k$ there is an $s_i \in \mathbb{F}_p$ such that $p_i(x)$ divides $g(x) - s_i$, i.e. $g(x) \equiv s_i \pmod{p_i(x)}$.

14. Prove that the polynomials $g(x)$ of degree $< n$ satisfying the equivalent conditions of the previous exercise form a vector space V over $\mathbb{F}_p$ of dimension k. [Use the Chinese Remainder Theorem applied to the p^k possible choices for the s_i in (d)].

15. Let $g(x) = b_0 + b_1 x + \dots + b_{n-1}x^{n-1} \in V$. For $j = 0, 1, \dots, n-1$ let

$$R(x^{pj}) = a_{0,j} + a_{1,j}x + \dots + a_{n-1,j}x^{n-1}$$

and let A be the $n \times n$ matrix

$$A = \begin{pmatrix} a_{0,0} & a_{0,1} & \dots & a_{0,n-1} \\ a_{1,0} & a_{1,1} & \dots & a_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1,0} & a_{n-1,1} & \dots & a_{n-1,n-1} \end{pmatrix}. \qquad (*)$$

Show that condition (a) of Exercise 13 for $g(x) \in V$ is equivalent to

$$(A - I)B = 0 \qquad (**)$$

where B is the column matrix with entries $b_0, b_1, \dots, b_{n-1}$. Conclude that the rank of the matrix $A - I$ is $n - k$. Note that this already suffices to determine if $f(x)$ is irreducible, without actually determining the factors.

16. Let $g_1(x), g_2(x), \dots, g_k(x)$ be a basis of solutions to $(**)$ (so a basis for V), where we may take $g_1(x) = 1$. Beginning with $w(x) = f(x)$, compute the greatest common divisor $(w(x), g_i(x) - s)$ for $i = 2, 3, \dots, k$ and $s \in \mathbb{F}_p$ for every factor of $f(x)$ already computed. Note by Exercise 13(d) that every factor $p_i(x)$ of $f(x)$ divides such a g.c.d. The process terminates when k relatively prime factors have been determined.

Prove that this procedure actually gives all the factors $p_1(x), p_2(x), \dots, p_k(x)$, i.e. one can separate the individual factors $p_1(x), p_2(x), \dots, p_k(x)$ by this procedure, as follows:

If this were not the case, then for two of the factors, say $p_1(x)$ and $p_2(x)$, for each $i = 1, 2, \dots, k$ there would exist $s_i \in \mathbb{F}_p$ such that $g_i(x) - s_i$ is divisible by both $p_1(x)$ and $p_2(x)$. By the Chinese Remainder Theorem, choose a $g(x) \in V$ satisfying $g(x) \equiv 0 \pmod{p_1(x)}$ and $g(x) \equiv 1 \pmod{p_2(x)}$. Write $g(x) = \sum_{i=1}^k c_i g_i(x)$ in terms of the basis for V and let $s = \sum_{i=1}^k c_i s_i(x) \in \mathbb{F}_p$. Show that $s \equiv 0 \pmod{p_1(x)}$ so that $s = 0$ and $s \equiv 1 \pmod{p_2(x)}$ so that $s = 1$, a contradiction.

17. This exercise follows Berlekamp's Factorization Algorithm outlined in the previous exercises to determine the factorization of $f(x) = x^5 + x^2 + 4x + 6$ in $\mathbb{F}_7[x]$.
(a) Show that $x^7 \equiv x^2 + 3x^3 + 6x^4 \pmod{f(x)}$. Similarly compute x^{14}, x^{21}, and x^{28} modulo $f(x)$ (note that x^{14} can most easily be computed by squaring the result for x^7 and then reducing, etc.) to show that in this case the matrix A in Exercise 15 is:

$$\begin{pmatrix} 1 & 0 & 5 & 1 & 4 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 1 & 3 & 3 & 3 \\ 0 & 3 & 4 & 2 & 2 \\ 0 & 6 & 3 & 1 & 1 \end{pmatrix}.$$

(b) Show that the reduced row echelon form for $A - I$ is the matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 6 \\ 0 & 0 & 1 & 0 & 6 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Conclude that $k = 2$ (so $f(x)$ is the product of precisely two factors which are powers of irreducible polynomials) and that $g_1(x) = 1$ and $g_2(x) = x^4 + 5x^3 + x^2 + x$ give a basis for the solutions to $(**)$ in Exercise 15.

(c) Following the procedure in Exercise 16, show that $(f(x), g_2(x) - 1) = x^2 + 3x + 5 = p_1(x)$, with $f(x)/p_1(x) = x^3 + 4x^2 + 4x + 4 = p_2(x)$, giving the powers of the irreducible polynomials dividing $f(x)$ in $\mathbb{F}_7[x]$. Show that neither factor is a 7th power in $\mathbb{F}_7[x]$ and that each is relatively prime to its derivative to conclude that both factors are irreducible polynomials, giving the complete factorization of $f(x)$ into irreducible polynomials:

$$f(x) = (x^2 + 3x + 5)(x^3 + 4x^2 + 4x + 4) \in \mathbb{F}_7[x].$$

14.4 COMPOSITE EXTENSIONS AND SIMPLE EXTENSIONS

We now consider the effect of taking composites with Galois extensions. The first result states that "sliding up" a Galois extension gives a Galois extension.

Proposition 19. Suppose K/F is a Galois extension and F'/F is any extension. Then KF'/F' is a Galois extension, with Galois group

$$\mathrm{Gal}(KF'/F') \cong \mathrm{Gal}(K/K \cap F')$$

isomorphic to a subgroup of $\mathrm{Gal}(K/F)$. Pictorially,

$$\begin{tikzcd}[column sep=small] & KF' \arrow[dl, dash] \arrow[dr, dash, "\|"] & \\ K \arrow[dr, dash, "\|"'] & & F' \arrow[dl, dash] \\ & K \cap F' \arrow[d, dash] & \\ & F & \end{tikzcd}$$

Proof: If K/F is Galois, then K is the splitting field of some separable polynomial $f(x)$ in $F[x]$. Then KF'/F' is the splitting field of $f(x)$ viewed as a polynomial in $F'[x]$, hence this extension is Galois. Since K/F is Galois, every embedding of K fixing F is an automorphism of K, so the map

$$\begin{aligned} \varphi : \mathrm{Gal}(KF'/F') &\to \mathrm{Gal}(K/F) \\ \sigma &\mapsto \sigma|_K \end{aligned}$$

defined by restricting an automorphism σ to the subfield K is well defined. It is clearly a homomorphism, with kernel

$$\ker \varphi = \{\sigma \in \mathrm{Gal}(KF'/F') \mid \sigma|_K = 1\}.$$

Since an element in $\mathrm{Gal}(KF'/F')$ is trivial on F', the elements in the kernel are trivial both on K and on F', hence on their composite, so the kernel consists only of the identity automorphism. Hence φ is injective.

Let H denote the image of φ in $\mathrm{Gal}(K/F)$ and let K_H denote the corresponding fixed subfield of K containing F. Since every element in H fixes F', K_H contains $K \cap F'$. On the other hand, the composite $K_H F'$ is fixed by $\mathrm{Gal}(KF'/F')$ (any $\sigma \in \mathrm{Gal}(KF'/F')$ fixes F' and acts on $K_H \subseteq K$ via its restriction $\sigma|_K \in H$, which fixes K_H by definition). By the Fundamental Theorem it follows that $K_H F' = F'$, so that $K_H \subseteq F'$, which gives the reverse inclusion $K_H \subseteq K \cap F'$. Hence $K_H = K \cap F'$, so again by the Fundamental Theorem, $H = \mathrm{Gal}(K/K \cap F')$, completing the proof.

Corollary 20. Suppose K/F is a Galois extension and F'/F is any finite extension. Then

$$[KF' : F] = \frac{[K : F][F' : F]}{[K \cap F' : F]}.$$

Proof: This follows by the proposition from the equality $[KF' : F'] = [K : K \cap F']$ given by the orders of the Galois groups in the proposition.

The example $F = \mathbb{Q}$, $K = \mathbb{Q}(\sqrt[3]{2})$, $F' = \mathbb{Q}(\rho\sqrt[3]{2})$, ρ a primitive 3rd root of unity, shows that the formula of Corollary 20 does not hold in general if neither of the two extensions is Galois.

Proposition 21. Let K_1 and K_2 be Galois extensions of a field F. Then

(1) The intersection $K_1 \cap K_2$ is Galois over F.

(2) The composite K_1K_2 is Galois over F. The Galois group is isomorphic to the subgroup

$$H = \{(\sigma, \tau) \mid \sigma|_{K_1 \cap K_2} = \tau|_{K_1 \cap K_2}\}$$

of the direct product $\mathrm{Gal}(K_1/F) \times \mathrm{Gal}(K_2/F)$ consisting of elements whose restrictions to the intersection $K_1 \cap K_2$ are equal.

$$\begin{tikzcd}
 & K_1K_2 \arrow[dl, dash] \arrow[dr, dash] & \\
K_1 \arrow[dr, dash] & & K_2 \arrow[dl, dash] \\
 & K_1 \cap K_2 \arrow[d, dash] & \\
 & F &
\end{tikzcd}$$

Proof: (1) Suppose $p(x)$ is an irreducible polynomial in $F[x]$ with a root α in $K_1 \cap K_2$. Since $\alpha \in K_1$ and K_1/F is Galois, all the roots of $p(x)$ lie in K_1. Similarly

all the roots lie in K_2, hence all the roots of $p(x)$ lie in $K_1 \cap K_2$. It follows easily that $K_1 \cap K_2$ is Galois as in Theorem 13.
(2) If K_1 is the splitting field of the separable polynomial $f_1(x)$ and K_2 is the splitting field of the separable polynomial $f_2(x)$ then the composite is the splitting field for the square free part of the polynomial $f_1(x)f_2(x)$, hence is Galois over F.

The map

$$\begin{aligned} \varphi : \mathrm{Gal}(K_1K_2/F) &\to \mathrm{Gal}(K_1/F) \times \mathrm{Gal}(K_2/F) \\ \sigma &\mapsto (\sigma|_{K_1}, \sigma|_{K_2}) \end{aligned}$$

is clearly a homomorphism. The kernel consists of the elements σ which are trivial on both K_1 and K_2, hence trivial on the composite, so the map is injective. The image lies in the subgroup H, since

$$(\sigma|_{K_1})|_{K_1 \cap K_2} = (\sigma|_{K_2})|_{K_1 \cap K_2} = \sigma|_{K_1 \cap K_2}.$$

The order of H can be computed by observing that for every $\sigma \in \mathrm{Gal}(K_1/F)$ there are $|\mathrm{Gal}(K_2/K_1 \cap K_2)|$ elements $\tau \in \mathrm{Gal}(K_2/F)$ whose restrictions to $K_1 \cap K_2$ are $\sigma|_{K_1 \cap K_2}$. Hence

$$\begin{aligned} |H| &= |\mathrm{Gal}(K_1/F)| \cdot |\mathrm{Gal}(K_2/K_1 \cap K_2)| \\ &= |\mathrm{Gal}(K_1/F)| \frac{|\mathrm{Gal}(K_2/F)|}{|\mathrm{Gal}(K_1 \cap K_2/F)|}. \end{aligned}$$

By Corollary 20 and the diagram above we see that the orders of H and $\mathrm{Gal}(K_1K_2/F)$ are then both equal to

$$[K_1K_2 : F] = \frac{[K_1 : F][K_2 : F]}{[K_1 \cap K_2 : F]}.$$

Hence the image of φ is precisely H, completing the proof.

Corollary 22. Let K_1 and K_2 be Galois extensions of a field F with $K_1 \cap K_2 = F$. Then

$$\mathrm{Gal}(K_1K_2/F) \cong \mathrm{Gal}(K_1/F) \times \mathrm{Gal}(K_2/F).$$

Conversely, if K is Galois over F and $G = \mathrm{Gal}(K/F) = G_1 \times G_2$ is the direct product of two subgroups G_1 and G_2, then K is the composite of two Galois extensions K_1 and K_2 of F with $K_1 \cap K_2 = F$.

Proof: The first part follows immediately from the proposition. For the second, let K_1 be the fixed field of $G_1 \subset G$ and let K_2 be the fixed field of $G_2 \subset G$. Then $K_1 \cap K_2$ is the field corresponding to the subgroup G_1G_2, which is all of G in this case, so $K_1 \cap K_2 = F$. The composite K_1K_2 is the field corresponding to the subgroup $G_1 \cap G_2$, which is the identity here, so $K_1K_2 = K$, completing the proof.

Corollary 23. Let E/F be any finite separable extension. Then E is contained in an extension K which is Galois over F and is minimal in the sense that any other Galois extension of F containing E contains K.

Proof: There exists a Galois extension of F containing E, for example the composite of the splitting fields of the minimal polynomials for a basis for E over F (which are all separable since E is separable over F). Then the intersection of all the Galois extensions of F containing E is our field K.

Definition. The Galois extension K of F containing E in the previous corollary is called the *Galois closure* of E over F.

It is often simpler to work in a Galois extension (for example in computing degrees as in Corollary 20). The existence of a Galois closure for a separable extension is frequently useful for reducing computations to consideration of Galois extensions.

Recall that an extension K of F is called *simple* if $K = F(\theta)$ for some element θ, called a *primitive element*.

Proposition 24. Let K/F be a finite extension. Then $K = F(\theta)$ if and only if there exist only finitely many subfields of K containing F.

Proof: Suppose first that $K = F(\theta)$ is simple. Let E be a subfield of K containing F:

$$\begin{array}{c} K \\ | \\ E \\ | \\ F \end{array}$$

Let $f(x) \in F[x]$ be the minimal polynomial for θ over F and let $g(x) \in E[x]$ be the minimal polynomial for θ over E. Then $g(x)$ divides $f(x)$ in $E[x]$. Let E' be the field generated over F by the coefficients of $g(x)$. Then $E' \subseteq E$ and clearly the minimal polynomial for θ over E' is still $g(x)$. But then

$$[K : E] = \deg g(x) = [K : E']$$

implies that $E = E'$. It follows that the subfields of K containing F are the subfields generated by the coefficients of the monic factors of $f(x)$, hence there are finitely many such subfields.

Suppose conversely that there are finitely many subfields of K containing F. If F is a finite field, then we have already seen that K is a simple extension (Proposition 17). Hence we may suppose F is infinite. It clearly suffices to show that $F(\alpha, \beta)$ is generated by a single element since K is finitely generated over F. Consider the subfields

$$F(\alpha + c\beta), \qquad c \in F.$$

Then since there are infinitely many choices for $c \in F$ and only finitely many such subfields, there exist c, c' in F, $c \neq c'$, with

$$F(\alpha + c\beta) = F(\alpha + c'\beta).$$

Then

$$\alpha + c\beta \in F(\alpha + c\beta)$$
$$\alpha + c'\beta \in F(\alpha + c\beta)$$

shows by taking the difference that $(c - c')\beta \in F(\alpha + c\beta)$, hence $\beta \in F(\alpha + c\beta)$ and then also $\alpha \in F(\alpha + c\beta)$. Hence

$$F(\alpha, \beta) \subseteq F(\alpha + c\beta)$$

and since the reverse inclusion is obvious, we have

$$F(\alpha, \beta) = F(\alpha + c\beta),$$

completing the proof.

Theorem 25. (The Primitive Element Theorem) If K/F is finite and separable, then K/F is simple. In particular any finite extension of fields of characteristic 0 is simple.

Proof: Let L be the Galois closure of K over F. Then any subfield of K containing F corresponds to a subgroup of the Galois group $\mathrm{Gal}(L/F)$ by the Fundamental Theorem. Since there are only finitely many such subgroups, the previous proposition shows that K/F is simple. The last statement follows since any finite extension of fields in characteristic 0 is separable.

As the proof of the proposition indicates, a primitive element for an extension can be obtained as a simple linear combination of the generators for the extension. In the case of Galois extensions it is only necessary to determine a linear combination which is not fixed by any nontrivial element of the Galois group since then by the Fundamental Theorem this linear combination could not lie in any proper subfield.

Examples

(1) The element $\sqrt{2} + \sqrt{3}$ generates the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ as we have already seen (it is not fixed by any of the four Galois automorphisms of this field).

(2) The field $\overline{\mathbb{F}_p}(x, y)$ of rational functions in the variables x and y over the algebraic closure $\overline{\mathbb{F}_p}$ of $\mathbb{F}_p$ is not a simple extension of the subfield $F = \overline{\mathbb{F}_p}(x^p, y^p)$. It is easy to see that

$$[\overline{\mathbb{F}_p}(x, y) : \overline{\mathbb{F}_p}(x^p, y^p)] = p^2$$

and that the subfields

$$F(x + cy), \qquad c \in \overline{\mathbb{F}_p}$$

are all of degree p over $\overline{\mathbb{F}_p}(x^p, y^p)$ (note that $(x + cy)^p = x^p + c^p y^p \in \overline{\mathbb{F}_p}(x^p, y^p)$). If any two of these subfields were equal, then just as in the proof of Proposition 24 we would have

$$\overline{\mathbb{F}_p}(x, y) = F(x + cy)$$

which is impossible by degree considerations. Hence there are infinitely many such subfields and the extension cannot be simple.

EXERCISES

1. Determine the Galois closure of the field $\mathbb{Q}(\sqrt{1+\sqrt{2}})$.
2. Find a primitive generator for $\mathbb{Q}(\sqrt{2}, \sqrt{3}, \sqrt{5})$.
3. Let F be a field contained in the ring of $n \times n$ matrices over $\mathbb{Q}$. Prove that $[F : \mathbb{Q}] \leq n$. (Note that, by Exercise 19 of Section 13.2, the ring of $n \times n$ matrices over $\mathbb{Q}$ does contain fields of degree n over $\mathbb{Q}$.)
4. Let $f(x) \in F[x]$ be an irreducible polynomial of degree n over the field F, let L be the splitting field of $f(x)$ over F and let α be a root of $f(x)$ in L. If K is any Galois extension of F, show that the polynomial $f(x)$ splits into a product of m irreducible polynomials each of degree d over K, where $d = [K(\alpha) : K] = [(L \cap K)(\alpha) : L \cap K]$ and $m = n/d = [F(\alpha) \cap K : F]$. [Show first that the factorization of $f(x)$ over K is the same as its factorization over $L \cap K$. Then if H is the subgroup of the Galois group of L over F corresponding to $L \cap K$ the factors of $f(x)$ over $L \cap K$ correspond to the orbits of H on the roots of $f(x)$. Use Exercise 9 of Section 4.1.]
5. Let p be a prime and let F be a field. Let K be a Galois extension of F whose Galois group is a p-group (i.e. the degree $[K : F]$ is a power of p). Such an extension is called a *p-extension* (note that p-extensions are Galois by definition).
   (a) Let L be a p-extension of K. Prove that the Galois closure of L over F is a p-extension of F.
   (b) Give an example to show that (a) need not hold if $[K : F]$ is a power of p but K/F is not Galois.
6. Prove that $\mathbb{F}_p(x, y)/\mathbb{F}_p(x^p, y^p)$ is not a simple extension by explicitly exhibiting an infinite number of intermediate subfields.

14.5 CYCLOTOMIC EXTENSIONS AND ABELIAN EXTENSIONS OVER $\mathbb{Q}$

We have already determined that the cyclotomic field $\mathbb{Q}(\zeta_n)$ of nth roots of unity is a Galois extension of $\mathbb{Q}$ of degree $\varphi(n)$ where φ denotes the Euler φ-function. Any automorphism of this field is uniquely determined by its action on the primitive n^{th} root of unity ζ_n. This element must be mapped to another primitive n^{th} root of unity (recall these are the roots of the irreducible cyclotomic polynomial $\Phi_n(x)$). Hence $\sigma(\zeta_n) = \zeta_n^a$ for some integer a, $1 \leq a < n$, relatively prime to n. Since there are precisely $\varphi(n)$ such integers a it follows that in fact each of these maps is indeed an automorphism of $\mathbb{Q}(\zeta_n)$. Note also that we can define σ_a for any integer a relatively prime to n by the same formula and that σ_a depends only on the residue class of a modulo n.

Theorem 26. The Galois group of the cyclotomic field $\mathbb{Q}(\zeta_n)$ of nth roots of unity is isomorphic to the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^\times$, the isomorphism being given by

$$\begin{aligned} (\mathbb{Z}/n\mathbb{Z})^\times &\xrightarrow{\sim} \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \\ a \pmod{n} &\mapsto \sigma_a \end{aligned}$$

where σ_a is the automorphism defined by

$$\sigma_a(\zeta_n) = \zeta_n^a.$$

Proof: The discussion above shows that σ_a is an automorphism for any $a \pmod n$, so the map above is well-defined. It is a homomorphism since

$$\begin{aligned}(\sigma_a\sigma_b)(\zeta_n) &= \sigma_a(\zeta_n^b) = (\zeta_n^b)^a \\ &= \zeta_n^{ab}\end{aligned}$$

which shows that $\sigma_a\sigma_b = \sigma_{ab}$. The map is bijective by the discussion above since we know that every Galois automorphism is of the form σ_a for a uniquely defined $a \pmod n$. Hence the map is an isomorphism.

Examples

(1) The field $\mathbb{Q}(\zeta_5)$ is Galois over $\mathbb{Q}$ with Galois group $(\mathbb{Z}/5\mathbb{Z})^\times \cong \mathbb{Z}/4\mathbb{Z}$. This is our first example of a Galois extension of $\mathbb{Q}$ of degree 4 with a *cyclic* Galois group. The elements of the Galois group are $\{\sigma_1 = 1, \sigma_2, \sigma_3, \sigma_4\}$ in the notation above. A generator for this cyclic group is $\sigma_2 : \zeta_5 \mapsto \zeta_5^2$ (since 2 has order 4 in $(\mathbb{Z}/5\mathbb{Z})^\times$).

There is precisely one nontrivial subfield, a quadratic extension of $\mathbb{Q}$, the fixed field of the subgroup $\{1, \sigma_4 = \sigma_{-1}\}$. An element in this subfield is given by

$$\alpha = \zeta_5 + \sigma_{-1}\zeta_5 = \zeta_5 + \zeta_5^{-1}$$

since this element is clearly fixed by σ_{-1}. The element ζ_5 satisfies

$$\zeta_5{}^4 + \zeta_5{}^3 + \zeta_5{}^2 + \zeta_5 + 1 = 0.$$

Notice then that

$$\begin{aligned}\alpha^2 + \alpha - 1 &= (\zeta_5{}^2 + 2 + \zeta_5{}^{-2}) + (\zeta_5 + \zeta_5^{-1}) - 1 \\ &= \zeta_5{}^2 + 2 + \zeta_5{}^3 + \zeta_5 + \zeta_5^4 - 1 = 0.\end{aligned}$$

Solving explicitly for α we see that the quadratic extension of $\mathbb{Q}$ generated by α is $\mathbb{Q}(\sqrt{5})$:

$$\mathbb{Q}(\zeta_5 + \zeta_5^{-1}) = \mathbb{Q}(\sqrt{5}).$$

It can be shown in general (this is not completely trivial) that for p an odd prime the field $\mathbb{Q}(\zeta_p)$ contains the quadratic field $\mathbb{Q}(\sqrt{\pm p})$, where the $+$ sign is correct if $p \equiv 1 \mod 4$ and the $-$ sign is correct if $p \equiv 3 \mod 4$.

(2) For p an odd prime we can construct a primitive element for any of the subfields of $\mathbb{Q}(\zeta_p)$ as in the previous example. A basis for $\mathbb{Q}(\zeta_p)$ over $\mathbb{Q}$ is given by

$$1, \zeta_p, \zeta_p^2, \ldots, \zeta_p^{p-2}.$$

Since

$$\zeta_p^{p-1} + \zeta_p^{p-2} + \ldots + \zeta_p + 1 = 0$$

we see that also the elements

$$\zeta_p, \zeta_p^2, \ldots, \zeta_p^{p-2}, \zeta_p^{p-1}$$

form a basis. The reason for choosing this basis is that any σ in the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_p)/\mathbb{Q})$ simply *permutes* these basis elements since these are precisely the primitive p^{th} roots of unity. Note that it is at this point that we need p to be a prime — in general the primitive n^{th} roots of unity do not give a basis for the cy-

clotomic field of nth roots of unity over $\mathbb{Q}$ (for example, the primitive 4th roots of unity, $\pm i$, are not linearly independent).

Let H be any subgroup of the Galois group of $\mathbb{Q}(\zeta_p)$ over $\mathbb{Q}$ and let

$$\alpha_H = \sum_{\sigma \in H} \sigma \zeta_p, \tag{14.10}$$

the sum of the conjugates of ζ_p by the elements in H. For any $\tau \in H$, the elements $\tau\sigma$ run over the elements of H as σ runs over the elements of H. It follows that $\tau\alpha = \alpha$, so that α lies in the fixed field for H. If now τ is *not* an element of H, then $\tau\alpha$ is the sum of basis elements (recall that any automorphism permutes the basis elements here), one of which is $\tau(\zeta_p)$. If we had $\tau\alpha = \alpha$ then since these elements are a *basis*, we must have $\tau(\zeta_p) = \sigma(\zeta_p)$ for one of the terms $\sigma\zeta_p$ in (10). But this implies $\tau\sigma^{-1} = 1$ since this automorphism is the identity on ζ_p. Then $\tau = \sigma \in H$, a contradiction. This shows that α is not fixed by any automorphism not contained in H, so that $\mathbb{Q}(\alpha)$ is precisely the fixed field of H.

For a specific example, consider the subfields of $\mathbb{Q}(\zeta_{13})$, which correspond to the subgroups of $(\mathbb{Z}/13\mathbb{Z})^\times \cong \mathbb{Z}/12\mathbb{Z}$. A generator for this cyclic group is the automorphism $\sigma = \sigma_2$ which maps ζ_{13} to ζ_{13}^2. The nontrivial subgroups correspond to the nontrivial divisors of 12, hence are of orders 2,3,4, and 6 with generators σ^6, σ^4, σ^3 and σ^2, respectively. The corresponding fixed fields will be of degrees 6,4,3 and 2 over $\mathbb{Q}$, respectively. Generators are given by ($\zeta = \zeta_{13}$)

$$\begin{aligned}
\zeta + \sigma^6\zeta &= \zeta + \zeta^{2^6} = \zeta + \zeta^{-1} \\
\zeta + \sigma^4\zeta + \sigma^8\zeta &= \zeta + \zeta^{2^4} + \zeta^{2^8} = \zeta + \zeta^3 + \zeta^9 \\
\zeta + \sigma^3\zeta + \sigma^6\zeta + \sigma^9\zeta &= \zeta + \zeta^8 + \zeta^{12} + \zeta^5 \\
\zeta + \sigma^2\zeta + \sigma^4\zeta + \sigma^6\zeta + \sigma^8\zeta + \sigma^{10}\zeta &= \zeta + \zeta^4 + \zeta^3 + \zeta^{12} + \zeta^9 + \zeta^{10}.
\end{aligned}$$

The lattice of subfields for this extension is

$$\begin{tikzcd}
 & \mathbb{Q}(\zeta) \arrow[dl, dash, "3"'] \arrow[d, dash, "2"] & \\
\mathbb{Q}(\zeta+\zeta^3+\zeta^9) \arrow[dr, dash, "2"'] & \mathbb{Q}(\zeta+\zeta^{-1}) \arrow[dr, dash, "2"] \arrow[dd, dash, "3"] & \\
 & & \mathbb{Q}(\zeta+\zeta^5+\zeta^8+\zeta^{12}) \arrow[d, dash, "3"] \\
 & \mathbb{Q}(\zeta+\zeta^3+\zeta^4+\zeta^9+\zeta^{10}+\zeta^{12}) \arrow[r, dash, "2"'] & \mathbb{Q}
\end{tikzcd}$$

The elements constructed in equation (10) and their conjugates are called the *periods* of ζ and are useful in the study of the arithmetic of the cyclotomic fields. The study of their combinatorial properties is referred to as *cyclotomy*.

Suppose that $n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$ is the decomposition of n into distinct prime powers. Since $\zeta_n^{p_2^{a_2} \dots p_k^{a_k}}$ is a primitive $p_1^{a_1}$-th root of unity, the field $K_1 =$

$\mathbb{Q}(\zeta_{p_1^{a_1}})$ is a subfield of $\mathbb{Q}(\zeta_n)$. Similarly, each of the fields $K_i = \mathbb{Q}(\zeta_{p_i^{a_i}})$, $i = 1, 2, \ldots, k$ is a subfield of $\mathbb{Q}(\zeta_n)$. The composite of the fields contains the product $\zeta_{p_1^{a_1}}\zeta_{p_2^{a_2}}\ldots\zeta_{p_k^{a_k}}$, which is a primitive n^{th} root of unity, hence the composite field is $\mathbb{Q}(\zeta_n)$. Since the extension degrees $[K_i : \mathbb{Q}]$ equal $\varphi(p_i^{a_i})$, $i = 1, 2, \ldots, k$ and $\varphi(n) = \varphi(p_1^{a_1})\varphi(p_2^{a_2})\ldots\varphi(p_k^{a_k})$, the degree of the composite of the fields K_i is precisely the product of the degrees of the K_i. It follows from Proposition 21 (and a simple induction from the two fields considered in the proposition to the k fields here) that the intersections of all these fields is precisely $\mathbb{Q}$. Then Corollary 22 shows that the Galois group for $\mathbb{Q}(\zeta_n)$ is the direct product of the Galois groups over $\mathbb{Q}$ for the subfields K_i. We summarize this as the following corollary.

Corollary 27. Let $n = p_1^{a_1}p_2^{a_2}\ldots p_k^{a_k}$ be the decomposition of the positive integer n into distinct prime powers. Then the cyclotomic fields $\mathbb{Q}(\zeta_{p_i^{a_i}})$, $i = 1, 2, \ldots, k$ intersect only in the field $\mathbb{Q}$ and their composite is the cyclotomic field $\mathbb{Q}(\zeta_n)$. We have

$$\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q}) \cong \mathrm{Gal}(\mathbb{Q}(\zeta_{p_1^{a_1}})/\mathbb{Q}) \times \mathrm{Gal}(\mathbb{Q}(\zeta_{p_2^{a_2}})/\mathbb{Q}) \times \cdots \times \mathrm{Gal}(\mathbb{Q}(\zeta_{p_k^{a_k}})/\mathbb{Q})$$

which under the isomorphism in Theorem 26 is the Chinese Remainder Theorem:

$$(\mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/p_1^{a_1}\mathbb{Z})^\times \times (\mathbb{Z}/p_2^{a_2}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_k^{a_k}\mathbb{Z})^\times .$$

Proof: The only statement which has not been proved is the identification of the isomorphism of Galois groups with the statement of the Chinese Remainder Theorem on the group $(\mathbb{Z}/n\mathbb{Z})^\times$, which is quite simple and is left for the exercises.

By Theorem 26 the Galois group of $\mathbb{Q}(\zeta_n)/\mathbb{Q}$ is in particular an abelian group.

Definition. The extension K/F is called an *abelian* extension if K/F is Galois and $\mathrm{Gal}(K/F)$ is an abelian group.

Since all the subgroups and quotient groups of abelian groups are abelian, we see by the Fundamental Theorem of Galois Theory that every subfield containing F of an abelian extension of F is again an abelian extension of F. By the results on composites of extensions in the last section, we also see that the composite of abelian extensions is again an abelian extension (since the Galois group of the composite is isomorphic to a subgroup of the direct product of the Galois groups, hence is abelian).

It is an open problem to determine which groups arise as the Galois groups of Galois extensions of $\mathbb{Q}$. Using the results above we can see that every *abelian* group appears as the Galois group of some extension of $\mathbb{Q}$, in fact as the Galois group of some subfield of a cyclotomic field.

Let $n = p_1p_2\ldots p_k$ be the product of distinct primes. Then by the Chinese Remainder Theorem

$$\begin{aligned}(\mathbb{Z}/n\mathbb{Z})^\times &\cong (\mathbb{Z}/p_1\mathbb{Z})^\times \times (\mathbb{Z}/p_2\mathbb{Z})^\times \times \ldots \times (\mathbb{Z}/p_k\mathbb{Z})^\times \\ &\cong Z_{p_1-1} \times Z_{p_2-1} \times \ldots \times Z_{p_k-1} .\end{aligned} \tag{14.11}$$

Now, suppose G is any finite abelian group. By the Fundamental Theorem for Abelian Groups,

$$G \cong Z_{n_1} \times Z_{n_2} \times \ldots \times Z_{n_k}$$

for some integers $n_1, n_2, \ldots, n_k$. We take as known that given any integer m there are infinitely many primes p with $p \equiv 1 \mod m$ (see the exercises following Section 13.6 for one proof using cyclotomic polynomials). Given this result, choose distinct primes $p_1, p_2, \ldots, p_k$ such that

$$\begin{aligned} p_1 &\equiv 1 \mod n_1 \\ p_2 &\equiv 1 \mod n_2 \\ &\vdots \\ p_k &\equiv 1 \mod n_k \end{aligned}$$

and let $n = p_1 p_2 \ldots p_k$ as above.

By construction, n_i divides $p_i - 1$ for $i = 1, 2, \ldots, k$, so the group $Z_{p_i - 1}$ has a subgroup H_i of order $\dfrac{p_i - 1}{n_i}$ for $i = 1, 2, \ldots, k$, and the quotient by this subgroup is cyclic of order n_i. Hence the quotient of $(\mathbb{Z}/n\mathbb{Z})^\times$ in equation (11) by $H_1 \times H_2 \times \ldots \times H_k$ is isomorphic to the group G.

By Theorem 26 and the Fundamental Theorem of Galois Theory, we see that there is a subfield of $\mathbb{Q}(\zeta_{p_1 p_2 \ldots p_k})$ which is Galois over $\mathbb{Q}$ with G as Galois group. We summarize this as:

Corollary 28. Let G be any finite abelian group. Then there is a subfield K of a cyclotomic field with $\mathrm{Gal}(K/\mathbb{Q}) \cong G$.

There is a converse to this result (whose proof is beyond our scope), the celebrated Kronecker–Weber Theorem:

Theorem. (Kronecker–Weber) Let K be a finite abelian extension of $\mathbb{Q}$. Then K is contained in a cyclotomic extension of $\mathbb{Q}$.

The abelian extensions of $\mathbb{Q}$ are the "easiest" Galois extensions (at least in so far as the structure of their Galois groups is concerned) and the previous result shows they can be classified by the cyclotomic extensions of $\mathbb{Q}$. For other finite extensions of $\mathbb{Q}$ as base field, it is more difficult to describe the abelian extensions. The study of the abelian extensions of an arbitrary finite extension F of $\mathbb{Q}$ is referred to as *class field theory*. There is a classification of the abelian extensions of F by invariants associated to F which greatly generalizes the results on cyclotomic fields over $\mathbb{Q}$. In general, however, the construction of abelian extensions is not nearly as explicit as in the case of the cyclotomic fields. One case where such a description is possible is for the abelian extensions of an imaginary quadratic field ($\mathbb{Q}(\sqrt{-D})$ for D positive), where the abelian extensions can be constructed by adjoining values of certain elliptic functions (this is the analogue of adjoining the roots of unity, which are the values of

the exponential function e^x for certain x). The study of the arithmetic of such abelian extensions and the search for similar results for non-abelian extensions are rich and fascinating areas of current mathematical research.

We end our discussion of the cyclotomic fields with the problem of the constructibility of the regular n-gon by straightedge and compass.

Recall (cf. Section 13.3) that an element α is constructible over $\mathbb{Q}$ if and only if the field $K = \mathbb{Q}(\alpha)$ is obtained by a series of quadratic extensions:

$$\mathbb{Q} = K_0 \subset K_1 \subset \ldots \subset K_i \subset K_{i+1} \subset \ldots \subset K_m = K \tag{14.12}$$

with

$$[K_{i+1} : K_i] = 2, \qquad i = 0, 1, \ldots, m-1.$$

The construction of the regular n-gon in $\mathbb{R}^2$ is evidently equivalent to the construction of the nth roots of unity, since the nth roots of unity form the vertices of a regular n-gon on the unit circle in $\mathbb{C}$ with one vertex at the point 1.

The construction of ζ_n is equivalent to the constructibility of the first coordinate x in $\mathbb{R}^2$ of ζ_n, namely the real part of ζ_n. Since the complex conjugate of ζ_n is just ζ_n^{-1}, the real part of ζ_n is $x = \frac{1}{2}(\zeta_n + \zeta_n^{-1})$. Note that ζ_n satisfies the quadratic equation $\zeta_n^2 - 2x\zeta_n + 1 = 0$ over $\mathbb{Q}(x)$. Since $\mathbb{Q}(x)$ consists only of real numbers, it follows that $[\mathbb{Q}(\zeta_n) : \mathbb{Q}(x)] = 2$, so that $\mathbb{Q}(x)$ is an extension of degree $\varphi(n)/2$ of $\mathbb{Q}$.

It follows that if the regular n-gon can be constructed by straightedge and compass then $\varphi(n)$ must be a power of 2. Conversely, if $\varphi(n) = 2^m$ is a power of 2, then the Galois group $\mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ is an abelian group whose order is a power of 2, so the same is true for the Galois group $\mathrm{Gal}(\mathbb{Q}(x)/\mathbb{Q})$. It is easy to see by the Fundamental Theorem for Abelian Groups that an abelian group G of order 2^m has a chain of subgroups

$$G = G_m > G_{m-1} > \ldots > G_{i+1} > G_i > \ldots > G_0 = 1$$

with

$$[G_{i+1} : G_i] = 2, \qquad i = 0, 1, 2, \ldots, m-1.$$

Applying this to the group $G = \mathrm{Gal}(\mathbb{Q}(x)/\mathbb{Q})$ and taking the fixed fields for the subgroups G_i, $i = 0, 1, \ldots, m-1$, we obtain (by the Fundamental Theorem of Galois Theory) a sequence of quadratic extensions as in (12) above.

We conclude that the regular n-gon can be constructed by straightedge and compass if and only if $\varphi(n)$ is a power of 2. Decomposing n into prime powers to compute $\varphi(n)$ we see that this means $n = 2^k p_1 \ldots p_r$ is the product of a power of 2 and distinct odd primes p_i where $p_i - 1$ is a power of 2. It is an elementary exercise to see that a prime p with $p - 1$ a power of 2 must be of the form

$$p = 2^{2^s} + 1$$

for some integer s. Such primes are called *Fermat primes*. The first few are

$$3 = 2^1 + 1$$
$$5 = 2^2 + 1$$

$$17 = 2^4 + 1$$
$$257 = 2^8 + 1$$
$$65537 = 2^{16} + 1$$

(but $2^{32} + 1$ is not a prime, being divisible by 641). It is not known if there are infinitely many Fermat primes. We summarize this as:

Proposition 29. The regular n-gon can be constructed by straightedge and compass if and only if $n = 2^k p_1 \ldots p_r$ is the product of a power of 2 and distinct Fermat primes.

The proof above actually indicates a procedure for constructing the regular n-gon as a succession of square roots. For example, the construction of the regular 17-gon (solved by Gauss in 1796 at age 19) requires the construction of the subfields of degrees 2,4,8 and 16 in $\mathbb{Q}(\zeta_{17})$. These subfields can be constructed by forming the *periods* of ζ_{17} as in the example of the 13th roots of unity above. In this case, the fact that $\mathbb{Q}(\zeta_{17})$ is obtained by a series of quadratic extensions reflects itself in the fact that the periods can be "halved" successively (i.e. if $H_1 < H_2$ are subgroups with $[H_2 : H_1] = 2$ then the periods for H_1 satisfy a quadratic equation whose coefficients involve the periods for H_2). For example, the periods for the subgroup of index 2 (generated by σ_2) in the Galois group are ($\zeta = \zeta_{17}$)

$$\eta_1 = \zeta + \zeta^2 + \zeta^4 + \zeta^8 + \zeta^9 + \zeta^{13} + \zeta^{15} + \zeta^{16}$$
$$\eta_2 = \zeta^3 + \zeta^5 + \zeta^6 + \zeta^7 + \zeta^{10} + \zeta^{11} + \zeta^{12} + \zeta^{14}$$

which "halve" the period for the full Galois group and which satisfy

$$\eta_1 + \eta_2 = -1$$

(from the minimal polynomial satisfied by ζ_{17}) and

$$\eta_1 \eta_2 = -4$$

(which requires computation — we know that it must be rational by Galois Theory, since this product is fixed by all the elements of the Galois group). Hence these two periods are the roots of the quadratic equation

$$x^2 + x - 4 = 0$$

which we can solve explicitly. In a similar way, the periods for the subgroup of index 4 (generated by σ_4) naturally halve these periods, so are quadratic over these, etc. In this way one can determine ζ_{17} explicitly in terms of iterated square roots. For example, one finds that $8(\zeta + \zeta^{-1}) = 16\cos(\frac{2\pi}{17})$ (which is enough to construct the regular 17-gon) is given explicitly by

$$-1 + \sqrt{17} + \sqrt{2(17 - \sqrt{17})} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{2(17 - \sqrt{17})} - 2\sqrt{2(17 + \sqrt{17})}}.$$

A relatively simple construction of the regular 17–gon (shown to us by J. Conway) is indicated in the exercises.

While we have seen that it is not possible to solve for ζ_n using only successive square roots in general, by definition it is possible to obtain ζ_n by successive extraction of higher roots (namely, taking an nth root of 1). This is not the case for solutions of general equations of degree n, where one cannot generally determine solutions by radicals, as we shall see in the next sections.

EXERCISES

1. Determine the minimal polynomials satisfied by the primitive generators given in the text for the subfields of $\mathbb{Q}(\zeta_{13})$.
2. Determine the subfields of $\mathbb{Q}(\zeta_8)$ generated by the periods of ζ_8 and in particular show that not every subfield has such a period as primitive element.
3. Determine the quadratic equation satisfied by the period $\alpha = \zeta_5 + \zeta_5^{-1}$ of the 5th root of unity ζ_5. Determine the quadratic equation satisfied by ζ_5 over $\mathbb{Q}(\alpha)$ and use this to explicitly solve for the 5th root of unity.
4. Let $\sigma_a \in \text{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ denote the automorphism of the cyclotomic field of nth roots of unity which maps ζ_n to ζ_n^a where a is relatively prime to n and ζ_n is a primitive n^{th} root of unity. Show that $\sigma_a(\zeta) = \zeta^a$ for *every* n^{th} root of unity.
5. Let p be a prime and let $\epsilon_1, \epsilon_2, \ldots, \epsilon_{p-1}$ denote the primitive p^{th} roots of unity. Set $p_n = \epsilon_1^n + \epsilon_2^n + \cdots + \epsilon_{p-1}^n$, the sum of the nth powers of the ϵ_i. Prove that $p_n = -1$ if p does not divide n and that $p_n = p - 1$ if p does divide n. [One approach: $p_1 = -1$ from $\Phi_p(x)$; show that p_n is a Galois conjugate of p_1 for p not dividing n, hence is also -1.]
6. Let ζ_n denote a primitive n^{th} root of unity and let $K = \mathbb{Q}(\zeta_n)$ be the associated cyclotomic field. Let a denote the trace of ζ_n from K to $\mathbb{Q}$ (cf. Exercise 18 of Section 2). Prove that $a = 1$ if $n = 1$, $a = 0$ if n is divisible by the square of a prime, and $a = (-1)^r$ if n is the product of r distinct primes.
7. Show that complex conjugation restricts to the automorphism $\sigma_{-1} \in \text{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ of the cyclotomic field of nth roots of unity. Show that the field $K^+ = \mathbb{Q}(\zeta_n + \zeta_n^{-1})$ is the subfield of real elements in $K = \mathbb{Q}(\zeta_n)$, called the *maximal real subfield of* K.
8. Let $K_n = \mathbb{Q}(\zeta_{2^{n+2}})$ be the cyclotomic field of 2^{n+2}-th roots of unity, $n \geq 0$. Set $\alpha_n = \zeta_{2^{n+2}} + \zeta_{2^{n+2}}^{-1}$ and $K_n^+ = \mathbb{Q}(\alpha_n)$, the maximal real subfield of K_n.
   (a) Show that for all $n \geq 0$, $[K_n : \mathbb{Q}] = 2^{n+1}$, $[K_n : K_n^+] = 2$, $[K_n^+ : \mathbb{Q}] = 2^n$, and $[K_{n+1}^+ : K_n^+] = 2$.
   (b) Determine the quadratic equation satisfied by $\zeta_{2^{n+2}}$ over K_n^+ in terms of α_n.
   (c) Show that for $n \geq 0$, $\alpha_{n+1}^2 = 2 + \alpha_n$ and hence show that
   $$\alpha_n = \sqrt{2 + \sqrt{2 + \sqrt{\cdots + \sqrt{2}}}} \qquad (n \textit{ times}),$$
   giving an explicit formula for the (constructible) 2^{n+2}-th roots of unity.
9. Notation as in the previous exercise.
   (a) Prove that K_n^+ is a cyclic extension of $\mathbb{Q}$ of degree 2^n. [Use an explicit isomorphism $(\mathbb{Z}/2^{n+2}\mathbb{Z})^\times \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2^n\mathbb{Z}$ as abelian groups (i.e. $(\mathbb{Z}/2^{n+2}\mathbb{Z})^\times$ is isomorphic to a cyclic group of order 2 and a cyclic group of order 2^n – cf. Exercises 22 and 23 of Section 2.3]
   (b) Prove that K_n is a biquadratic extension of K_{n-1}^+ and that two of the three intermediate subfields are K_n^+ and K_{n-1}. Prove that the remaining field intermediate between K_{n-1}^+ and K_n is a cyclic extension of $\mathbb{Q}$ of degree 2^n.

10. Prove that $\mathbb{Q}(\sqrt[3]{2})$ is not a subfield of any cyclotomic field over $\mathbb{Q}$.

11. Prove that the primitive n^{th} roots of unity form a basis over $\mathbb{Q}$ for the cyclotomic field of nth roots of unity if and only if n is square-free (i.e. n is not divisible by the square of any prime).

12. Let σ_p denote the Frobenius automorphism $x \mapsto x^p$ of the finite field $\mathbb{F}_q$ of $q = p^n$ elements. Viewing $\mathbb{F}_q$ as a vector space V of dimension n over $\mathbb{F}_p$ we can consider σ_p as a linear transformation of V to V. Determine the characteristic polynomial of σ_p and prove that σ_p is diagonalizable as a linear transformation if and only if n divides $q-1 = p^n - 1$.

13. Let $n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$ be the prime factorization of n and let ζ_n be a primitive n^{th} root of unity. For each i, $i = 1, 2, \dots, k$ define d_i by $n = p_i^{a_i} d_i$ and let $\zeta_{p_i^{a_i}} = \zeta_n^{d_i}$, so that $\zeta_{p_i^{a_i}}$ is a particular primitive $p_i^{a_i}$-th root of unity. Let $\sigma_a \in \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$ be the automorphism mapping ζ_n to ζ_n^a for a relatively prime to n.

(a) Prove that for $i = 1, 2, \dots, k$, σ_a maps $\zeta_{p_i^{a_i}}$ to $\zeta_{p_i^{a_i}}^a$ and gives an automorphism of $\mathbb{Q}(\zeta_{p_i^{a_i}})/\mathbb{Q}$ which depends only on $a \pmod{p_i^{a_i}}$, which we may denote $\sigma_{a \pmod{p_i^{a_i}}}$.

(b) Prove that the map $\sigma_a \mapsto (\sigma_{a \pmod{p_1^{a_1}}}, \sigma_{a \pmod{p_2^{a_2}}}, \dots, \sigma_{a \pmod{p_k^{a_k}}})$ is the isomorphism of Corollary 27 corresponding to the Chinese Remainder Theorem for $(\mathbb{Z}/n\mathbb{Z})^\times$.

The following Exercises 14 to 18 determine the periods associated to a primitive 17th root of unity and provide a proof for the simple geometric construction indicated in Exercise 17 for the regular 17–gon. Let $\zeta = \zeta_{17} = \cos\dfrac{2\pi}{17} + i \sin\dfrac{2\pi}{17}$ be a fixed primitive 17th root of unity in $\mathbb{C}$.

14. Define the *periods* of ζ as follows:

$$\begin{aligned}
\eta_1 &= \zeta + \zeta^2 + \zeta^4 + \zeta^8 + \zeta^9 + \zeta^{13} + \zeta^{15} + \zeta^{16} \\
\eta_2 &= \zeta^3 + \zeta^5 + \zeta^6 + \zeta^7 + \zeta^{10} + \zeta^{11} + \zeta^{12} + \zeta^{14} \\
\eta_1' &= \zeta + \zeta^4 + \zeta^{13} + \zeta^{16} \\
\eta_2' &= \zeta^2 + \zeta^8 + \zeta^9 + \zeta^{15} \\
\eta_3' &= \zeta^6 + \zeta^7 + \zeta^{10} + \zeta^{11} \\
\eta_4' &= \zeta^3 + \zeta^5 + \zeta^{12} + \zeta^{14} \\
\eta_1'' &= \zeta + \zeta^{16} \\
\eta_2'' &= \zeta^4 + \zeta^{13}.
\end{aligned}$$

(a) Show that all of these periods are real numbers and that $\eta_1'' = 2\cos\dfrac{2\pi}{17}$. Show that as real numbers these periods are approximately

$$\begin{aligned}
\eta_1 &\sim 1.562 \\
\eta_2 &\sim -2.562 \\
\eta_1' &\sim 2.049 \\
\eta_2' &\sim -0.488 \\
\eta_3' &\sim -2.906 \\
\eta_4' &\sim 0.344 \\
\eta_1'' &\sim 1.865 \\
\eta_2'' &\sim 0.185.
\end{aligned}$$

(b) Prove that η_1 and η_2 are roots of the equation $x^2 + x - 4 = 0$.

(c) Prove that η_1' and η_2' are roots of the equation $x^2 - \eta_1 x - 1 = 0$ and that η_3' and η_4' are roots of the equation $x^2 - \eta_2 x - 1 = 0$.

(d) Prove that η_1'' and η_2'' are roots of the equation $x^2 - \eta_1' x + \eta_4' = 0$.

15. Prove that if $\tan 2\theta = a$ $(0 < 2\theta < \frac{\pi}{2})$ then $\tan\theta$ satisfies the equation $x^2 - \frac{2}{a}x - 1 = 0$.

16. Let C be the circle in $\mathbb{R}^2$ having the points (h, k) and $(0, 1)$ as a diameter. Prove that this circle intersects the x–axis if and only if $h^2 - 4k \geq 0$ and in this case the two intercepts are the roots of the equation $x^2 - hx + k = 0$.

17. (*Construction of the regular 17–gon*) Draw a circle of radius 2 centered at the origin $(0, 0)$.

(a) Join the point $(4, 0)$ to the point $(0, 1)$ and construct the line ℓ_1 bisecting the angle between this line and the y–axis. Construct the line ℓ_2 perpendicular to ℓ_1 (Fig. 2).

(b) Using the intersection of ℓ_1 and the x–axis as center and radius equal to the distance to $(0, 1)$, construct the circle C_1 and let $A = (s, 0)$ be the right–hand point of intersection

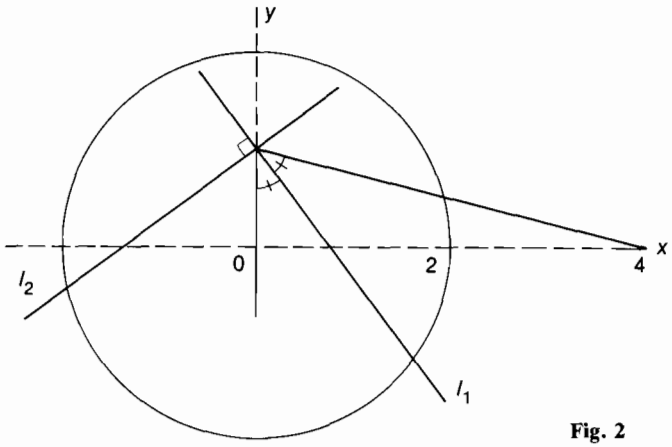


Fig. 2

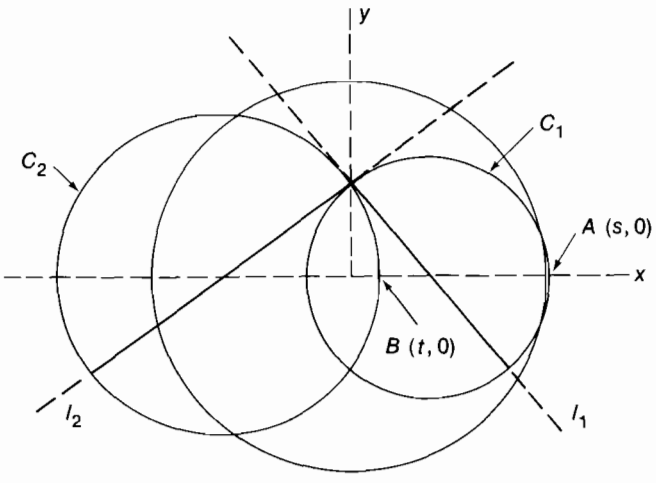


Fig. 3

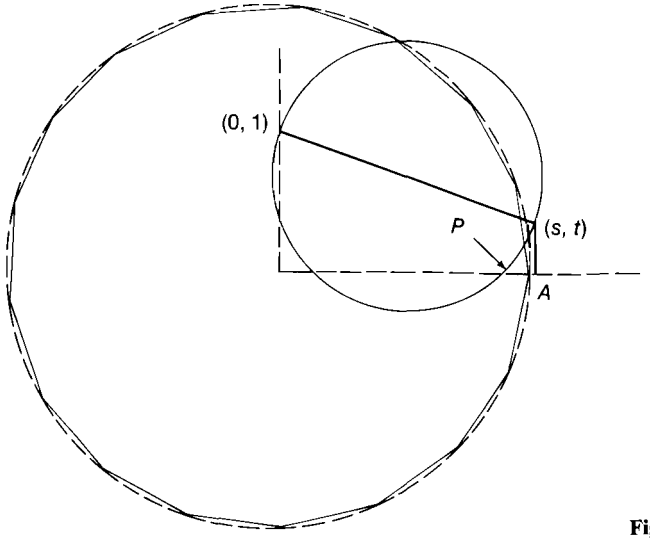


Fig. 4

of C_1 with the x–axis. Similarly, let $B = (t, 0)$ denote the right–hand point of intersection of the x–axis and the circle C_2 whose center is the intersection of ℓ_2 and the x–axis and whose radius is equal to the distance to $(0, 1)$ (Fig. 3).

(c) Construct a perpendicular to the x–axis at the point A and mark off the distance t from $(0, 0)$ to B to construct the point (s, t). Construct the circle with (s, t) and $(0, 1)$ as a diameter and let P denote the right–hand point of intersection of this circle with the x–axis. The perpendicular to the x–axis at P intersects the circle of radius 2 at the second vertex of a regular 17–gon whose first vertex is at (2,0), hence constructs the regular 17–gon by straightedge and compass (Fig. 4).

18. Notation as in the previous exercises.

(a) Prove that ℓ_1 intersects the x–axis in the point $(\eta_1/2, 0)$ and that ℓ_2 intersects the x–axis in the point $(\eta_2/2, 0)$.

(b) Prove that C_1 is the circle having the points $(\eta_1, -1)$ and $(0, 1)$ as diameter. Prove that $s = \eta_1'$. Similarly prove that C_2 is the circle having the points $(\eta_2, -1)$ and $(0, 1)$ as diameter and that $t = \eta_4'$.

(c) Prove that P has coordinates $(\eta_1'', 0)$ and hence that the construction in the previous problem constructs the regular 17–gon by straightedge and compass.

14.6 GALOIS GROUPS OF POLYNOMIALS

Recall that the Galois group of a separable polynomial $f(x) \in F[x]$ is defined to be the Galois group of the splitting field of $f(x)$ over F.

If K is a Galois extension of F then K is the splitting field for some separable polynomial $f(x)$ over F. Any automorphism $\sigma \in \mathrm{Gal}(K/F)$ maps a root of an irreducible factor of $f(x)$ to another root of the irreducible factor and σ is uniquely determined by its action on these roots (since they generate K over F). If we fix a labelling of the roots $\alpha_1, ..., \alpha_n$ of $f(x)$ we see that any $\sigma \in \mathrm{Gal}(K/F)$ defines a unique permutation of $\alpha_1, ..., \alpha_n$, hence defines a unique permutation of the subscripts $\{1, 2, ..., n\}$

(which depends on the fixed labelling of the roots). This gives us an injection

$$\mathrm{Gal}(K/F) \hookrightarrow S_n$$

of the Galois group into the symmetric group on n letters which is clearly a homomorphism (both group operations are composition). We may therefore think of Galois groups as subgroups of symmetric groups. Since the degree of the splitting field is the same as the order of the Galois group by the Fundamental Theorem, this explains from the group-theoretic side why the splitting field for a polynomial of degree n over F is of degree at most $n!$ over F (Proposition 13.26).

In general, if the factorization of $f(x)$ into irreducibles is $f(x) = f_1(x) \dots f_k(x)$ where $f_i(x)$ has degree n_i, $i = 1, 2, \dots, k$, then since the Galois group permutes the roots of the irreducible factors amongst themselves we have $\mathrm{Gal}(K/F) \leq S_{n_1} \times \dots \times S_{n_k}$.

If $f(x)$ is irreducible, then given any two roots of $f(x)$ there is an automorphism in the Galois group G of $f(x)$ which maps the first root to the second (this follows from our extension Theorem 13.27). Such a group is said to be *transitive* on the roots, i.e. you can get from any given root to any other root by applying some element of G. The fact that the Galois group must be transitive on blocks of roots (namely, the roots of the irreducible factors) can often be helpful in reducing the number of possibilities for the structure of G (cf. the discussion of Galois groups of polynomials of degree 4 below).

Examples

(1) Consider the biquadratic extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$, which is the splitting field of $(x^2 - 2)(x^2 - 3)$. Label the roots as $\alpha_1 = \sqrt{2}$, $\alpha_2 = -\sqrt{2}$, $\alpha_3 = \sqrt{3}$ and $\alpha_4 = -\sqrt{3}$. The elements of the Galois group are $\{1, \sigma, \tau, \sigma\tau\}$ where σ maps $\sqrt{2}$ to $-\sqrt{2}$ and fixes $\sqrt{3}$ and τ fixes $\sqrt{2}$ and maps $\sqrt{3}$ to $-\sqrt{3}$. As permutations of the roots for this labelling we see that σ interchanges the first two and fixes the second two and τ fixes the first two and interchanges the second two, i.e.

$$\sigma = (12) \qquad \text{and} \qquad \tau = (34)$$

as elements of S_4. Similarly, or by taking the product of these two elements, we see that

$$\sigma\tau = (12)(34) \in S_4.$$

Hence

$$\mathrm{Gal}(\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}) \cong \{1, (12), (34), (12)(34)\} \subset S_4$$

identifying this Galois group with the Klein-4 subgroup of S_4. Note that if we had changed the labelling of the roots above we would have obtained a different (isomorphic) representation of the Galois group as a subgroup of S_4 (for example, interchanging the second and third roots would have given the subgroup $\{1, (13), (24), (13)(24)\}$).

(2) The Galois group of $x^3 - 2$ acts as permutations on the three roots $\sqrt[3]{2}$, $\rho\sqrt[3]{2}$ and $\rho^2\sqrt[3]{2}$ where ρ is a primitive 3rd root of unity. With this ordering, the generators σ and τ we have defined earlier give the permutations

$$\sigma = (123) \qquad \tau = (23)$$

which gives

$$\{1, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\} = \{1, (123), (132), (23), (13), (12)\} = S_3,$$

in this case the full symmetric group on 3 letters.

Recall that every finite group is isomorphic to a subgroup of some symmetric group S_n. It is an open problem to determine whether every finite group appears as the Galois group for some polynomial over $\mathbb{Q}$. We have seen in the last section that every abelian group is a Galois group over $\mathbb{Q}$ (for some subfield of a cyclotomic field). We shall explicitly determine the Galois groups for polynomials of small degree (≤ 4) below which will in particular show that every subgroup of S_4 arises as a Galois group.

We first introduce some definitions and show that the "general" polynomial of degree n has S_n as Galois group (so the second example above should be viewed as "typical").

Definition. Let $x_1, x_2, ..., x_n$ be indeterminants. The *elementary symmetric functions* $s_1, s_2, ..., s_n$ are defined by

$$\begin{aligned} s_1 &= x_1 + x_2 + ... + x_n \\ s_2 &= x_1x_2 + x_1x_3 + ... + x_2x_3 + x_2x_4 + ... + x_{n-1}x_n \\ &\vdots \\ s_n &= x_1x_2...x_n \end{aligned}$$

i.e. the ith symmetric function s_i of $x_1, x_2, ..., x_n$ is the sum of all products of the x_j's taken i at a time.

Definition. The *general polynomial of degree* n is the polynomial

$$(x - x_1)(x - x_2)...(x - x_n)$$

whose roots are the indeterminants $x_1, x_2, ..., x_n$.

It is easy to see by induction that the coefficients of the general polynomial of degree n are given by the elementary symmetric functions in the roots:

$$(x - x_1)(x - x_2)...(x - x_n) = x^n - s_1x^{n-1} + s_2x^{n-2} + ... + (-1)^n s_n. \qquad (14.13)$$

For any field F, the extension $F(x_1, x_2, ..., x_n)$ is then a Galois extension of the field $F(s_1, s_2, ..., s_n)$ since it is the splitting field of the general polynomial of degree n.

If $\sigma \in S_n$ is any permutation of $\{1, 2, ..., n\}$, then σ acts on the rational functions in $F(x_1, x_2, ..., x_n)$ by permuting the subscripts of the variables $x_1, x_2, ..., x_n$. It is clear that this gives an automorphism of $F(x_1, x_2, ..., x_n)$. Identifying $\sigma \in S_n$ with this automorphism of $F(x_1, x_2, ..., x_n)$ identifies S_n as a subgroup of $\text{Aut}(F(x_1, x_2, ..., x_n))$.

The elementary symmetric functions $s_1, s_2, ..., s_n$ are fixed under any permutation of their subscripts (this is the reason they are called *symmetric*), which shows that the subfield $F(s_1, s_2, ..., s_n)$ is contained in the fixed field of S_n. By the Fundamental Theorem of Galois Theory, the fixed field of S_n has index precisely $n!$ in

$F(x_1, x_2, ..., x_n)$. Since $F(x_1, x_2, ..., x_n)$ is the splitting field over $F(s_1, s_2, ..., s_n)$ of the polynomial of degree n in (13), we have

$$[F(x_1, x_2, ..., x_n) : F(s_1, s_2, ..., s_n)] \leq n! \,. \tag{14.14}$$

It follows that we actually have equality and that $F(s_1, s_2, ..., s_n)$ is precisely the fixed field of S_n. This proves:

Proposition 30. The fixed field of the symmetric group S_n acting on the field of rational functions in n variables $F(x_1, x_2, ..., x_n)$ is the field of rational functions in the elementary symmetric functions $F(s_1, s_2, ..., s_n)$.

Definition. A rational function $f(x_1, x_2, ..., x_n)$ is called *symmetric* if it is not changed by any permutation of the variables $x_1, x_2, ..., x_n$.

Corollary 31. (Fundamental Theorem on Symmetric Functions) Any symmetric function in the variables $x_1, x_2, ..., x_n$ is a rational function in the elementary symmetric functions $s_1, s_2, ..., s_n$.

Proof: A symmetric function lies in the fixed field of S_n above, hence is a rational function in $s_1, ..., s_n$.

This corollary explains why these are called the *elementary* symmetric functions.

Remark: If $f(x_1, \ldots, x_n)$ is a *polynomial* in $x_1, x_2, ..., x_n$ which is symmetric then it can be seen that f is actually a polynomial in $s_1, s_2, ..., s_n$, which strengthens the statement of the corollary. It is in fact true that a symmetric polynomial whose coefficients lie in R, where R is any commutative ring with identity, is a polynomial in the elementary symmetric functions with coefficients in R. A proof of this fact is implicit in the algorithm outlined in the exercises for writing a symmetric polynomial as a polynomial in the elementary symmetric functions.

Examples

(1) The expression $(x_1 - x_2)^2$ is symmetric in x_1, x_2. We have

$$(x_1 - x_2)^2 = (x_1 + x_2)^2 - 4x_1x_2 = s_1^2 - 4s_2,$$

a polynomial in the elementary symmetric functions.

(2) The polynomial $x_1^2 + x_2^2 + x_3^2$ is symmetric in x_1, x_2, x_3, and in this case we have

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 &= (x_1 + x_2 + x_3)^2 - 2(x_1x_2 + x_1x_3 + x_2x_3) \\ &= s_1^2 - 2s_2. \end{aligned}$$

(3) The polynomial $x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2$ is symmetric. Since

$$\begin{aligned} (x_1x_2 + x_1x_3 + x_2x_3)^2 &= x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2 + 2(x_1^2x_2x_3 + x_2^2x_1x_3 + x_3^2x_1x_2) \\ &= x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2 + 2x_1x_2x_3(x_1 + x_2 + x_3) \end{aligned}$$

we have

$$x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2 = s_2^2 - 2s_1s_3.$$

Suppose now we *start* with the general polynomial

$$x^n - s_1x^{n-1} + s_2x^{n-2} + \ldots + (-1)^n s_n$$

over the field $F(s_1, s_2, \ldots, s_n)$ where we view the s_i, $i = 1, 2, \ldots, n$ as indeterminants. If we define the roots of this polynomial to be $x_1, x_2, \ldots, x_n$ then the s_i are precisely the elementary symmetric functions in the roots $x_1, \ldots, x_n$. From this point of view our result can be stated as:

Theorem 32. The general equation

$$x^n - s_1x^{n-1} + s_2x^{n-2} + \ldots + (-1)^n s_n$$

over the field $F(s_1, s_2, \ldots, s_n)$ is separable with Galois group S_n.

This result says that if there are no relations among the coefficients of a polynomial of degree n (which is what we mean when we say the s_i are indeterminants above) then the Galois group of this polynomial over the field generated by its coefficients is the full symmetric group S_n. Loosely speaking, this means that the "generic" polynomial of degree n will have S_n as Galois group. Note, however, that over finite fields every polynomial has a *cyclic* Galois group (all extensions of finite fields are cyclic), so that "generic" polynomials in this sense do not exist. Over $\mathbb{Q}$ one can make precise the notion of "generic" polynomial and then it is true that most polynomials have the full symmetric group as Galois group.

For $n \geq 5$ there is only one normal subgroup of S_n, namely the subgroup A_n of index 2. Hence in general there is only one normal subfield of $F(x_1, x_2, \ldots, x_n)$ containing $F(s_1, s_2, \ldots, s_n)$ and it is an extension of degree 2.

Definition. Define the *discriminant* D of $x_1, x_2, \ldots, x_n$ by the formula

$$D = \prod_{i<j}(x_i - x_j)^2.$$

Define the discriminant of a polynomial to be the discriminant of the roots of the polynomial.

The discriminant D is a symmetric function in $x_1, \ldots, x_n$, hence is an element of $K = F(s_1, s_2, \ldots, s_n)$.

When we first defined the alternating group A_n we saw that a permutation $\sigma \in S_n$ is an element of the subgroup A_n if and only if σ fixes the product

$$\sqrt{D} = \prod_{i<j}(x_i - x_j).$$

It follows (by the Fundamental Theorem) that $\sqrt{D}$ generates the fixed field of A_n and generates a quadratic extension of K. We state this as:

Proposition 33. The permutation $\sigma \in S_n$ is an element of A_n if and only if it fixes the square root of the discriminant D.

We now consider the Galois groups of separable polynomials of small degree (≤ 4) over a field F which we assume is of characteristic different from 2 and 3. Note that over $\mathbb{Q}$ or over a finite field (or, more generally, over any perfect field) the splitting field of an arbitrary polynomial $f(x)$ is the same as the splitting field for the product of the irreducible factors of $f(x)$ taken precisely once, which is a separable polynomial.

If the roots of the polynomial $f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0$ are $\alpha_1, \alpha_2, \ldots, \alpha_n$, then the discriminant of $f(x)$ is[2]

$$D = \prod_{i<j}(\alpha_i - \alpha_j)^2.$$

Note that $D = 0$ if and only if $f(x)$ is not separable, i.e. if the roots $\alpha_1, \ldots, \alpha_n$ are not distinct. Recall that over a perfect field (e.g. $\mathbb{Q}$ or a finite field) this implies $f(x)$ is reducible since every irreducible polynomial over a perfect field is separable.

The discriminant D is symmetric in the roots of $f(x)$, hence is fixed by all the automorphisms of the Galois group of $f(x)$. By the Fundamental Theorem it follows that $D \in F$. The discriminant can in general be written as a polynomial in the coefficients of $f(x)$ (by Corollary 31) which are fairly complicated for larger degrees (we shall give formulas for $n \leq 4$ below). Finally, note that since

$$\sqrt{D} = \prod_{i<j}(\alpha_i - \alpha_j)$$

we have the useful fact that $\sqrt{D}$ is always contained in the splitting field for $f(x)$.

If the roots of $f(x)$ are distinct, fix some ordering of the roots and view the Galois group of $f(x)$ as a subgroup of S_n as above.

Proposition 34. The Galois group of $f(x) \in F[x]$ is a subgroup of A_n if and only if the discriminant $D \in F$ is the square of an element of F.

Proof: This is a restatement of Proposition 33 in this case. The Galois group is contained in A_n if and only if every element of the Galois group fixes

$$\sqrt{D} = \prod_{i<j}(\alpha_i - \alpha_j)$$

i.e. if and only if $\sqrt{D} \in F$.

This property, together with the fact that $D = 0$ determines the presence of multiple roots, is the reason D is called the *discriminant*.

Polynomials of degree 2

Suppose the polynomial is $x^2 + ax + b$ with roots α, β. The discriminant D for this polynomial is $(\alpha - \beta)^2$, which can be written as a polynomial in the elementary symmetric

[2] If $f(x) = a_nx^n + \cdots + a_0$ is not monic then its discriminant is defined to be a_n^{2n-2} times the D defined above.

functions of the roots. We did this in Example 1 above:

$$D = s_1^2 - 4s_2 = (-a)^2 - 4(b) = a^2 - 4b,$$

the usual discriminant for this quadratic.

The polynomial is separable if and only if $a^2 - 4b \neq 0$. The Galois group is a subgroup of S_2, the cyclic group of order 2 and is trivial (i.e. A_2 in this case) if and only if $a^2 - 4b$ is a rational square, which completely determines the possible Galois groups.

Note that this restates results we obtained previously by explicitly solving for the roots: if the polynomial is reducible (namely D is a square in F), then the Galois group is trivial (the splitting field is just F), while if the polynomial is irreducible the Galois group is isomorphic to $\mathbb{Z}/2\mathbb{Z}$ since the splitting field is the quadratic extension $F(\sqrt{D})$.

Polynomials of degree 3

Suppose the cubic polynomial is

$$f(x) = x^3 + ax^2 + bx + c. \tag{14.15}$$

If we make the substitution $x = y - a/3$ the polynomial becomes

$$g(y) = y^3 + py + q \tag{14.16}$$

where

$$p = \frac{1}{3}(3b - a^2) \qquad q = \frac{1}{27}(2a^3 - 9ab + 27c)\,. \tag{14.17}$$

The splitting fields for these two polynomials are the same since their roots differ by the constant $a/3 \in F$ and since the formula for the discriminant involves the *differences* of roots, we see that these two polynomials also have the *same* discriminant.

Let the roots of the polynomial in (16) be α, β, and γ. We first compute the discriminant of this polynomial in terms of p and q. Note that

$$g(y) = (y - \alpha)(y - \beta)(y - \gamma)$$

so that if we differentiate we have

$$D_y g(y) = (y - \alpha)(y - \beta) + (y - \alpha)(y - \gamma) + (y - \beta)(y - \gamma).$$

Then

$$\begin{aligned} D_y g(\alpha) &= (\alpha - \beta)(\alpha - \gamma) \\ D_y g(\beta) &= (\beta - \alpha)(\beta - \gamma) \\ D_y g(\gamma) &= (\gamma - \alpha)(\gamma - \beta). \end{aligned}$$

Taking the product we see that

$$D = [(\alpha - \beta)(\alpha - \gamma)(\beta - \gamma)]^2 = -D_y g(\alpha) D_y g(\beta) D_y g(\gamma).$$

Since $D_y g(y) = 3y^2 + p$, we have

$$\begin{aligned} -D &= (3\alpha^2 + p)(3\beta^2 + p)(3\gamma^2 + p) \\ &= 27\alpha^2\beta^2\gamma^2 + 9p(\alpha^2\beta^2 + \alpha^2\gamma^2 + \beta^2\gamma^2) + 3p^2(\alpha^2 + \beta^2 + \gamma^2) + p^3. \end{aligned}$$

The corresponding expressions in the elementary symmetric functions of the roots were determined in Examples 2 and 3 above. Note that here $s_1 = 0$, $s_2 = p$ and $s_3 = -q$. We obtain

$$-D = 27(-q)^2 + 9p(p^2) + 3p^2(-2p) + p^3$$

so that

$$D = -4p^3 - 27q^2. \tag{14.18}$$

This is the same as the discriminant of $f(x)$ in (15). Expressing D in terms of a, b, c using (17) we obtain

$$D = a^2b^2 - 4b^3 - 4a^3c - 27c^2 + 18abc \tag{14.18'}$$

(Galois group of a cubic)

a. If the cubic polynomial $f(x)$ is reducible, then it splits either into three linear factors or into a linear factor and an irreducible quadratic. In the first case the Galois group is trivial and in the second case the Galois group is of order 2.

b. If the cubic polynomial $f(x)$ is irreducible then a root of $f(x)$ generates an extension of degree 3 over F, so the degree of the splitting field over F is divisible by 3. Since the Galois group is a subgroup of S_3, there are only two possibilities, namely A_3 and S_3. The Galois group is A_3 (i.e. cyclic of order 3) if and only if the discriminant D in (18) is a square.

Explicitly, if D is the square of an element of F, then the splitting field of the irreducible cubic $f(x)$ is obtained by adjoining any single root of $f(x)$ to F. The resulting field is Galois over F of degree 3 with a cyclic group of order 3 as Galois group. If D is not the square of an element of F then the splitting field of $f(x)$ is of degree 6 over F, hence is the field $F(\theta, \sqrt{D})$ for any one of the roots θ of $f(x)$. This extension is Galois over F with Galois group S_3 (generators are given by σ, which takes θ to one of the other roots of $f(x)$ and fixes $\sqrt{D}$, and τ, which takes $\sqrt{D}$ to $-\sqrt{D}$ and fixes θ).

We see that in both cases the splitting field for the irreducible cubic $f(x)$ is obtained by adjoining $\sqrt{D}$ and a root of $f(x)$ to F.

We shall give explicit formulas for the roots of (16) (*Cardano's formulas*) in the next section after introducing the notion of a *Lagrange Resolvent*.

Polynomials of degree 4

Let the quartic polynomial be

$$f(x) = x^4 + ax^3 + bx^2 + cx + d$$

which under the substitution $x = y - a/4$ becomes the quartic

$$g(y) = y^4 + py^2 + qy + r$$

with

$$p = \frac{1}{8}(-3a^2 + 8b)$$
$$q = \frac{1}{8}(a^3 - 4ab + 8c)$$
$$r = \frac{1}{256}(-3a^4 + 16a^2b - 64ac + 256d).$$

Let the roots of $g(y)$ be $\alpha_1, \alpha_2, \alpha_3$, and α_4 and let G denote the Galois group for the splitting field of $g(y)$ (or of $f(x)$).

Suppose first that $g(y)$ is reducible. If $g(y)$ splits into a linear and a cubic, then G is the Galois group of the cubic, which we determined above. Suppose then that $g(y)$ splits into two irreducible quadratics. Then the splitting field is the extension $F(\sqrt{D_1}, \sqrt{D_2})$ where D_1 and D_2 are the discriminants of the two quadratics. If D_1 and D_2 do not differ by a square then this extension is a biquadratic extension and G is isomorphic to the Klein 4-subgroup of S_4. If D_1 is a square times D_2 then this extension is a quadratic extension and G is isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

We are reduced to the situation where $g(y)$ is irreducible. In this case recall that the Galois group is transitive on the roots, i.e. it is possible to get from a given root to any other root by applying some automorphism of the Galois group. Examining the possibilities we see that the only transitive subgroups of S_4, hence the only possibilities for our Galois group G, are the groups:

$S_4, \quad A_4$

$D_8 = \{1, (1324), (12)(34), (1423), (13)(24), (14)(23), (12), (34)\}$ and its conjugates

$V = \{1, (12)(34), (13)(24), (14)(23)\}$

$C = \{1, (1234), (13)(24), (1432)\}$ and its conjugates.

(D_8 is the dihedral group, a Sylow 2-subgroup of S_4, with 3 (isomorphic) conjugate subgroups in S_4, V is the Klein 4-subgroup of S_4, normal in S_4, and C is a cyclic group, with 3 (isomorphic) conjugates in S_4).

Consider the elements

$$\theta_1 = (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)$$
$$\theta_2 = (\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4)$$
$$\theta_3 = (\alpha_1 + \alpha_4)(\alpha_2 + \alpha_3)$$

in the splitting field for $g(y)$. These elements are permuted amongst themselves by the permutations in S_4. The stabilizer of θ_1 in S_4 is the dihedral group D_8. The stabilizers in S_4 of θ_2 and θ_3 are the conjugate dihedral subgroups of order 8. The subgroup of S_4 which stabilizes all three of these elements is the intersection of these subgroups, namely the Klein 4-group V.

Since S_4 merely permutes $\theta_1, \theta_2, \theta_3$ it follows that the elementary symmetric functions in the θ's are fixed by all the elements of S_4, hence are in F. An elementary computation in symmetric functions shows that these elementary symmetric functions

are $2p, p^2 - 4r$, and $-q^2$, which shows that $\theta_1, \theta_2, \theta_3$ are the roots of

$$h(x) = x^3 - 2px^2 + (p^2 - 4r)x + q^2$$

called the *resolvent cubic* for the quartic $g(y)$. Since

$$\begin{aligned}\theta_1 - \theta_2 &= \alpha_1\alpha_3 + \alpha_2\alpha_4 - \alpha_1\alpha_2 - \alpha_3\alpha_4 \\ &= -(\alpha_1 - \alpha_4)(\alpha_2 - \alpha_3)\end{aligned}$$

and similarly

$$\begin{aligned}\theta_1 - \theta_3 &= -(\alpha_1 - \alpha_3)(\alpha_2 - \alpha_4) \\ \theta_2 - \theta_3 &= -(\alpha_1 - \alpha_2)(\alpha_3 - \alpha_4)\end{aligned}$$

we see that the discriminant of the resolvent cubic is the *same* as the discriminant of the quartic $g(y)$, hence also as the discriminant of the quartic $f(x)$. Using our formula for the discriminant of the cubic, we can easily compute the discriminant in terms of p, q, r:

$$D = 16p^4r - 4p^3q^2 - 128p^2r^2 + 144pq^2r - 27q^4 + 256r^3$$

from which one can give the formula for D in terms of a, b, c, d:

$$\begin{aligned}D = &- 128b^2d^2 - 4a^3c^3 + 16b^4d - 4b^3c^2 - 27a^4d^2 + 18abc^3 + 144a^2bd^2 - 192acd^2 \\ &+ a^2b^2c^2 - 4a^2b^3d - 6a^2c^2d + 144bc^2d + 256d^3 - 27c^4 - 80ab^2cd + 18a^3bcd\,.\end{aligned}$$

The splitting field for the resolvent cubic is a subfield of the splitting field of the quartic, so the Galois group of the resolvent cubic is a quotient of G. Hence knowing the action of the Galois group on the roots of the resolvent cubic $h(x)$ gives information about the Galois group of $g(y)$:

(Galois group of a quartic)

a. Suppose first that the resolvent cubic is irreducible. If D is not a square, then G is not contained in A_4 and the Galois group of the resolvent cubic is S_3, which implies that the degree of the splitting field for $g(y)$ is divisible by 6. The only possibility is then $G = S_4$.

b. If the resolvent cubic is irreducible and D is a square, then G is a subgroup of A_4 and 3 divides the order of G (the Galois group of the resolvent cubic is A_3). The only possibility is $G = A_4$.

c1. We are left with the case where the resolvent cubic is reducible. The first possibility is that $h(x)$ has 3 roots in F (i.e. splits completely). Since each of the elements $\theta_1, \theta_2, \theta_3$ is in F, every element of G fixes all three of these elements, which means $G \subseteq V$. The only possibility is $G = V$.

c2. If $h(x)$ splits into a linear and a quadratic, then precisely one of $\theta_1, \theta_2, \theta_3$ is in F, say θ_1. Then G stabilizes θ_1 but not θ_2 and θ_3, so we have $G \subseteq D_8$ and $G \not\subseteq V$. This leaves two possibilities: $G = D_8$ and $G = C$. One way to distinguish between these is to observe that $F(\sqrt{D})$ is the fixed field of the elements of G in A_4. For the two cases being considered, we have $D_8 \cap A_4 = V$, $C \cap A_4 = \{1, (13)(24)\}$. The first

group is transitive on the roots of $g(y)$, the second is not. It follows that the first case occurs if and only if $g(y)$ is irreducible over $F(\sqrt{D})$. We may therefore determine G completely by factoring $g(y)$ in $F(\sqrt{D})$, and so completely determine the Galois group in all cases. (cf. the exercises following and in the next section, where it is shown that over $\mathbb{Q}$ the Galois group cannot be cyclic of degree 4 if D is not the sum of two squares – so in particular if $D < 0$.)

We shall give explicit formulas for the roots of a quartic polynomial at the end of the next section.

The Fundamental Theorem of Algebra

We end this section with two proofs of the Fundamental Theorem of Algebra. We need two facts regarding the field $\mathbb{C}$:

(a) Every polynomial with real coefficients of odd degree has a root in the reals. Equivalently, there are no nontrivial finite extensions of $\mathbb{R}$ of odd degree.

(b) Quadratic polynomials with coefficients in $\mathbb{C}$ have roots in $\mathbb{C}$. Equivalently, there are no quadratic extensions of $\mathbb{C}$.

The first result follows from the Intermediate Value Theorem in calculus, since the graph of a monic polynomial $f(x) \in \mathbb{R}[x]$ of odd degree is negative for large negative values of x and positive for large positive values of x, hence crosses the axis somewhere. The equivalence with the second statement follows since a finite extension of $\mathbb{R}$ is a simple extension and the minimal polynomial of a primitive element would have odd degree, hence would be both irreducible over $\mathbb{R}$ and have a root in $\mathbb{R}$, hence must be of degree 1.

The second result follows by a direct computation. By the quadratic formula it suffices to show that every complex number $\alpha = a + bi$, $a, b \in \mathbb{R}$, has a square root in $\mathbb{C}$. Write $\alpha = re^{i\theta}$ for some $r \geq 0$ and some $\theta \in [0, 2\pi)$. Then $\sqrt{r}e^{i\theta/2}$ is a square root of α. (Explicitly, let $c \in \mathbb{R}$ be a square root of the real number $\dfrac{a + \sqrt{a^2 + b^2}}{2}$ and let $d \in \mathbb{R}$ be a square root of the real number $\dfrac{-a + \sqrt{a^2 + b^2}}{2}$ where the signs of the two square roots are chosen so that cd has the same sign as b. Then multiplying out we see that $(c + di)^2 = a + bi$.)

Theorem 35. (Fundamental Theorem of Algebra) Every polynomial $f(x) \in \mathbb{C}[x]$ of degree n has precisely n roots in $\mathbb{C}$ (counted with multiplicity). Equivalently, $\mathbb{C}$ is algebraically closed.

Proof: I. It suffices to prove that every polynomial $f(x) \in \mathbb{C}[x]$ has a root in $\mathbb{C}$. Let τ denote the automorphism complex conjugation. If $f(x)$ has no root in $\mathbb{C}$ then neither does the conjugate polynomial $\bar{f}(x) = \tau f(x)$ obtained by applying τ to the coefficients of $f(x)$ (since its roots are the conjugates of the roots of $f(x)$). The product $f(x)\bar{f}(x)$ has coefficients which are invariant under complex conjugation, hence

has real coefficients. It suffices then to prove that a polynomial with real coefficients has a root in $\mathbb{C}$.

Suppose that $f(x)$ is a polynomial of degree n with real coefficients and write $n = 2^k m$ where m is odd. We prove that $f(x)$ has a root in $\mathbb{C}$ by induction on k. For $k = 0$, $f(x)$ has odd degree and by (a) above $f(x)$ has a root in $\mathbb{R}$ so we are done. Suppose now that $k \geq 1$. Let $\alpha_1, \alpha_2, ..., \alpha_n$ be the roots of $f(x)$ and set $K = \mathbb{R}(\alpha_1, \alpha_2, ..., \alpha_n, i)$. Then K is a Galois extension of $\mathbb{R}$ containing $\mathbb{C}$ and the roots of $f(x)$. For any $t \in \mathbb{R}$ consider the polynomial

$$L_t = \prod_{1 \leq i < j \leq n} [x - (\alpha_i + \alpha_j + t\alpha_i\alpha_j)].$$

Any automorphism of $K/\mathbb{R}$ permutes the terms in this product so the coefficients of L_t are invariant under all the elements of $\mathrm{Gal}(K/\mathbb{R})$. Hence L_t is a polynomial with real coefficients. The degree of L_t is

$$\frac{n(n-1)}{2} = 2^{k-1}m(2^k m - 1) = 2^{k-1}m'$$

where m' is odd (since $k \geq 1$). The power of 2 in this degree is therefore less than k, so by induction the polynomial L_t has a root in $\mathbb{C}$. Hence for each $t \in \mathbb{R}$ one of the elements $\alpha_i + \alpha_j + t\alpha_i\alpha_j$ for some i, j $(1 \leq i < j \leq n)$ is an element of $\mathbb{C}$. Since there are infinitely many choices for t and only finitely many values of i and j we see that for some i and j (say, $i = 1$ and $j = 2$) there are distinct real numbers s and t with

$$\alpha_1 + \alpha_2 + s\alpha_1\alpha_2 \in \mathbb{C} \qquad \alpha_1 + \alpha_2 + t\alpha_1\alpha_2 \in \mathbb{C}.$$

Since $s \neq t$ it follows that $a = \alpha_1 + \alpha_2 \in \mathbb{C}$ and $b = \alpha_1\alpha_2 \in \mathbb{C}$. But then α_1 and α_2 are the roots of the quadratic $x^2 - ax + b$ with coefficients in $\mathbb{C}$, hence are elements of $\mathbb{C}$ by (b) above, completing the proof.

II. The second proof again uses (a) and (b) above, but replaces the computations with the polynomials L_t above with a simple group-theoretic argument involving the nilpotency of a Sylow 2-subgroup of the Galois group:

Let $f(x)$ be a polynomial of degree n with real coefficients and let K be the splitting field of $f(x)$ over $\mathbb{R}$. Then $K(i)$ is a Galois extension of $\mathbb{R}$. Let G denote its Galois group and let P_2 denote a Sylow 2-subgroup of G. The fixed field of P_2 is an extension of $\mathbb{R}$ of odd degree, hence by (a) is trivial.

It follows that $\mathrm{Gal}(K(i)/\mathbb{C})$ is a 2-group. Since 2-groups have subgroups of all orders (recall this is true of a finite p-group for any prime p, cf. Theorem 6.1), if this group is nontrivial, there would exist a quadratic extension of $\mathbb{C}$, impossible by (b), completing the proof.

The Fundamental Theorem of Algebra was first rigorously proved by Gauss in 1816 (his doctoral dissertation in 1798 provides a proof using geometric considerations requiring some topological justification). The first proof above is essentially due to Laplace in 1795 (hence the reason for naming the polynomials L_t). The reason Laplace's proof was deemed unacceptable was that he assumed the existence of a split-

ting field for polynomials (i.e. that the roots existed *somewhere* in *some* field), which had not been established at that time. The elegant second proof is a simplification due to Artin.

EXERCISES

1. Show that a cubic with a multiple root has a linear factor. Is the same true for quartics?
2. Determine the Galois groups of the following polynomials:
   (a) $x^3 - x^2 - 4$
   (b) $x^3 - 2x + 4$
   (c) $x^3 - x + 1$
   (d) $x^3 + x^2 - 2x - 1$.
3. Prove for any $a, b \in \mathbb{F}_{p^n}$ that if $x^3 + ax + b$ is irreducible then $-4a^3 - 27b^2$ is a square in $\mathbb{F}_{p^n}$.
4. Determine the Galois group of $x^4 - 25$.
5. Determine the Galois group of $x^4 + 4$.
6. Determine the Galois group of $x^4 + 3x^3 - 3x - 2$.
7. Determine the Galois group of $x^4 + 2x^2 + x + 3$.
8. Determine the Galois group of $x^4 + 8x + 12$.
9. Determine the Galois group of $x^4 + 4x - 1$ (cf. Exercise 19).
10. Determine the Galois group of $x^5 + x - 1$.
11. Let F be an extension of $\mathbb{Q}$ of degree 4 that is not Galois over $\mathbb{Q}$. Prove that the Galois closure of F has Galois group either S_4, A_4 or the dihedral group D_8 of order 8. Prove that the Galois group is dihedral if and only if F contains a quadratic extension of $\mathbb{Q}$.
12. Prove that an extension F of $\mathbb{Q}$ of degree 4 can be generated by the root of an irreducible biquadratic $x^4 + ax^2 + b$ over $\mathbb{Q}$ if and only if F contains a quadratic extension of $\mathbb{Q}$.
13. **(a)** Let $\pm\alpha, \pm\beta$ denote the roots of the polynomial $f(x) = x^4 + ax^2 + b \in \mathbb{Z}[x]$. Prove that $f(x)$ is irreducible if and only if $\alpha^2, \alpha \pm \beta$ are not elements of $\mathbb{Q}$.[3]
    (b) Suppose $f(x)$ is irreducible and let G be the Galois group of $f(x)$. Prove that
    (i) $G \cong V$, the Klein 4-group, if and only if b is a square in $\mathbb{Q}$ if and only if $\alpha\beta \in \mathbb{Q}$ is rational.
    (ii) $G \cong C$, the cyclic group of order 4, if and only if $b(a^2 - 4b)$ is a square in $\mathbb{Q}$ if and only if $\mathbb{Q}(\alpha\beta) = \mathbb{Q}(\alpha^2)$.
    (iii) $G \cong D_8$, the dihedral group of order 8, if and only if b and $b(a^2 - 4b)$ are not squares in $\mathbb{Q}$ if and only if $\alpha\beta \notin \mathbb{Q}(\alpha^2)$.
14. Prove the polynomial $x^4 - px^2 + q \in \mathbb{Q}[x]$ is irreducible for any distinct odd primes p and q and has as Galois group the dihedral group of order 8.[4]
15. Prove the polynomial $x^4 + px + p \in \mathbb{Q}[x]$ is irreducible for every prime p and for $p \neq 3, 5$ has Galois group S_4. Prove the Galois group for $p = 3$ is dihedral of order 8 and for $p = 5$ is cyclic of order 4.[5]

---

[3] cf. the note An Elementary Test for the Galois Group of a Quartic Polynomial, Luise-Charlotte Kappe and Bette Warren, *Amer. Math. Monthly*, **96** (1989), pp. 133–137.

[4] ibid.

[5] ibid.

16. Determine the Galois group over $\mathbb{Q}$ of the polynomial x^4+8x^2+8x+4. Determine which of the subfields of this field are Galois over $\mathbb{Q}$ and for those which are Galois determine a polynomial $f(x) \in \mathbb{Q}[x]$ for which they are the splitting field over $\mathbb{Q}$.

17. Find the Galois group of $x^4 - 7$ over $\mathbb{Q}$ explicitly as a permutation group on the roots.

18. Let θ be a root of $x^3 - 3x + 1$. Prove that the splitting field of this polynomial is $\mathbb{Q}(\theta)$ and that the Galois group is cyclic of order 3. In particular the other roots of this polynomial can be written in the form $a + b\theta + c\theta^2$ for some $a, b, c \in \mathbb{Q}$. Determine the other roots explicitly in terms of θ.

19. Let $f(x)$ be an irreducible polynomial of degree 4 in $\mathbb{Q}[x]$ with discriminant D. Let K denote the splitting field of $f(x)$, viewed as a subfield of the complex numbers $\mathbb{C}$.
    (a) Prove that $\mathbb{Q}(\sqrt{D}) \subset K$.
    (b) Let τ denote complex conjugation and let τ_K denote the restriction of complex conjugation to K. Prove that τ_K is an element of $\mathrm{Gal}(K/\mathbb{Q})$ of order 1 or 2 depending on whether every element of K is real or not.
    (c) Prove that if $D < 0$ then K cannot be cyclic of degree 4 over $\mathbb{Q}$ (i.e. $\mathrm{Gal}(K/\mathbb{Q})$ cannot be a cyclic group of order 4).
    (d) Prove generally that $\mathbb{Q}(\sqrt{D})$ for squarefree $D < 0$ is not a subfield of a cyclic quartic field (cf. also Exercise 19 of Section 7).

20. Determine the Galois group of $(x^3-2)(x^3-3)$ over $\mathbb{Q}$. Determine all the subfields which contain $\mathbb{Q}(\rho)$ where ρ is a primitive 3rd root of unity.

21. Let $G \leq S_n$ be a subgroup of the symmetric group and suppose $\sigma_1, \ldots, \sigma_k$ are generators for G. Show that if the function $f(x_1, x_2, \ldots, x_n)$ is fixed by the generators σ_i then it is fixed by G.

22. (*Newton's Formulas*) Let $f(x)$ be a monic polynomial of degree n with roots $\alpha_1, \ldots, \alpha_n$. Let s_i be the elementary symmetric function of degree i in the roots and define $s_i = 0$ for $i > n$. Let $p_i = \alpha_1^i + \cdots + \alpha_n^i$, $i \geq 0$, be the sum of the ith powers of the roots of $f(x)$. Prove *Newton's formulas*:

$$
\begin{aligned}
&p_1 - s_1 = 0\\
&p_2 - s_1p_1 + 2s_2 = 0\\
&p_3 - s_1p_2 + s_2p_1 - 3s_3 = 0\\
&\quad\vdots\\
&p_i - s_1p_{i-1} + s_2p_{i-2} - \cdots + (-1)^{i-1}s_{i-1}p_1 + (-1)^i i s_i = 0
\end{aligned}
$$

23. (a) If $x+y+z=1$, $x^2+y^2+z^2=2$ and $x^3+y^3+z^3=3$, determine $x^4+y^4+z^4$.
    (b) Prove generally that x, y, z are not rational but that $x^n+y^n+z^n$ is rational for every positive integer n.

24. Prove that an $n \times n$ matrix A over a field of characteristic 0 is nilpotent if and only if the trace of A^k is 0 for all $k \geq 0$.

25. Prove that two $n \times n$ matrices A and B over a field of characteristic 0 have the same characteristic polynomial if and only if the trace of A^k equals the trace of B^k for all $k \geq 0$.

26. Use the fact that the trace of AB is the same as the trace of BA for any two $n \times n$ matrices A and B to show that AB and BA have the same characteristic polynomial over a field of characteristic 0 (the same result is true over a field of arbitrary characteristic).

27. Let $f(x)$ be a monic polynomial of degree n with roots $\alpha_1, \alpha_2, \ldots, \alpha_n$.

(a) Show that the discriminant D of $f(x)$ is the square of the Vandermonde determinant

$$\begin{vmatrix} 1 & \alpha_1 & \alpha_1^2 & \ldots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \ldots & \alpha_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \ldots & \alpha_n^{n-1} \end{vmatrix} = \prod_{i>j}(\alpha_i - \alpha_j).$$

(b) Taking the Vandermonde matrix above, multiplying on the left by its transpose and taking the determinant show that one obtains

$$D = \begin{vmatrix} p_0 & p_1 & p_2 & \ldots & p_{n-1} \\ p_1 & p_2 & p_3 & \ldots & p_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n-1} & p_n & p_{n+1} & \ldots & p_{2n-2} \end{vmatrix}$$

where $p_i = \alpha_1^i + \cdots + \alpha_n^i$ is the sum of the ith powers of the roots of $f(x)$, which can be computed in terms of the coefficients of $f(x)$ using Newton's formulas above. This gives an efficient procedure for calculating the discriminant of a polynomial.

28. Let α be a root of the irreducible polynomial $f(x) \in F[x]$ and let $K = F(\alpha)$. Let D be the discriminant of $f(x)$. Prove that $D = (-1)^{n(n-1)/2}\mathrm{N}_{K/F}(f'(\alpha))$, where $f'(x) = D_x f(x)$ is the derivative of $f(x)$.

The following exercises describe the *resultant* of two polynomials and in particular provide another efficient method for calculating the discriminant of a polynomial.

29. Let F be a field and let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ and $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$ be two polynomials in $F[x]$.

(a) Prove that a necessary and sufficient condition for $f(x)$ and $g(x)$ to have a common root (or, equivalently, a common divisor in $F[x]$) is the existence of a polynomial $a(x) \in F[x]$ of degree at most $m-1$ and a polynomial $b(x) \in F[x]$ of degree at most $n-1$ with $a(x)f(x) = b(x)g(x)$.

(b) Writing $a(x)$ and $b(x)$ explicitly as polynomials show that equating coefficients in the equation $a(x)f(x) = b(x)g(x)$ gives a system of $n+m$ linear equations for the coefficients of $a(x)$ and $b(x)$. Prove that this system has a nontrivial solution (hence $f(x)$ and $g(x)$ have a common zero) if and only if the determinant

$$R(f,g) = \begin{vmatrix} a_n & a_{n-1} & \ldots & a_0 & & & & \\ & a_n & a_{n-1} & \ldots & a_0 & & & \\ & & a_n & a_{n-1} & \ldots & a_0 & & \\ & & & \ddots & & & & \\ & & & & a_n & a_{n-1} & \ldots & a_0 \\ b_m & b_{m-1} & \ldots & b_0 & & & & \\ & b_m & b_{m-1} & \ldots & b_0 & & & \\ & & b_m & b_{m-1} & \ldots & b_0 & & \\ & & & \ddots & & & & \\ & & & & b_m & b_{m-1} & \ldots & b_0 \end{vmatrix}$$

is nonzero. Here $R(f,g)$, called the *resultant* of the two polynomials, is the determinant of an $(n+m)\times(n+m)$ matrix R with m rows involving the coefficients of $f(x)$ and n rows involving the coefficients of $g(x)$.

30. **(a)** With notations as in the previous problem, show that we have the matrix equation

$$R\begin{pmatrix} x^{n+m-1} \\ x^{n+m-2} \\ \vdots \\ x \\ 1 \end{pmatrix} = \begin{pmatrix} x^{m-1}f(x) \\ x^{m-2}f(x) \\ \vdots \\ f(x) \\ x^{n-1}g(x) \\ x^{n-2}g(x) \\ \vdots \\ g(x) \end{pmatrix}.$$

(b) Let R' denote the matrix of cofactors of R as in Theorem 11.27, so $R'R = R(f,g)I$, where I is the identity matrix. Multiply both sides of the matrix equation above by R' and equate the bottom entry of the resulting column matrices to prove that there are polynomials $r(x), s(x) \in F[x]$ with $r(x)f(x) + s(x)g(x) = R(f,g)$, i.e. the resultant of two polynomials is a linear combination (in $F[x]$) of the polynomials.

31. Consider $f(x)$ and $g(x)$ as general polynomials and suppose the roots of $f(x)$ are $x_1, x_2, \ldots, x_n$ and the roots of $g(x)$ are $y_1, y_2, \ldots, y_m$. The coefficients of $f(x)$ are powers of a_0 times the elementary symmetric functions in $x_1, x_2, \ldots, x_n$ and the coefficients of $g(x)$ are powers of b_0 times the elementary symmetric functions in $y_1, y_2, \ldots, y_m$.

(a) By expanding the determinant show that $R(f,g)$ is homogeneous of degree m in the coefficients a_i and homogeneous of degree n in the coefficients b_j.

(b) Show that $R(f,g)$ is $a_0^m b_0^n$ times a symmetric function in $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$.

(c) Since $R(f,g)$ is 0 if $f(x)$ and $g(x)$ have a common root, say $x_i = y_j$, show that $R(f,g)$ is divisible by $x_i - y_j$ for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Conclude by degree considerations that

$$R(f,g) = a_0^m b_0^n \prod_{i=1}^{n}\prod_{j=1}^{m}(x_i - y_j).$$

(d) Show that the product in (c) can be also be written

$$R(f,g) = a_0^m \prod_{i=1}^{n} g(x_i) = (-1)^{nm} b_0^n \prod_{j=1}^{m} f(y_j).$$

This gives an interesting *reciprocity* between the product of g evaluated at the roots of f and the product of f evaluated at the roots of g.

32. Consider now the special case where $g(x) = f'(x)$ is the derivative of $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ and suppose the roots of $f(x)$ are $\alpha_1, \alpha_2, \ldots, \alpha_n$. Using the formula

$$R(f,f') = \prod_{i=1}^{n} f'(\alpha_i)$$

of the previous exercise, prove that

$$D = (-1)^{n(n-1)/2} R(f, f')$$

where D is the discriminant of $f(x)$.

33. **(a)** Prove that the discriminant of the cyclotomic polynomial $\Phi_p(x)$ of the pth roots of unity for an odd prime p is $(-1)^{(p-1)/2} p^{p-2}$ [One approach: use Exercise 5 of the previous section together with the determinant form for the discriminant in terms of the power sums p_i.]

(b) Prove that $\mathbb{Q}(\sqrt{(-1)^{(p-1)/2} p}) \subset \mathbb{Q}(\zeta_p)$ for p an odd prime. (cf. also Exercise 11 of Section 7)

34. Use the previous exercise to prove that every quadratic extension of $\mathbb{Q}$ is contained in a cyclotomic extension (a special case of the Kronecker–Weber Theorem).

35. Prove that the discriminant D of the polynomial $x^n + px + q$ is $(-1)^{n(n-1)/2} n^n q^{n-1} + (-1)^{(n-1)(n-2)/2} (n-1)^{n-1} p^n$.

36. Prove that the discriminant of the polynomial $x^n + nx^{n-1} + n(n-1)x^{n-2} + \cdots + n(n-1)\ldots(3)(2)x + n!$ is $(-1)^{n(n-1)/2}(n!)^n$.

The following problems 37 to 43 outline two procedures for writing a symmetric function in terms of the elementary symmetric functions. Let $f(x_1, \ldots, x_n)$ be a polynomial which is symmetric in $x_1, \ldots, x_n$. Define the *weight* of the monomial term $Ax_1^{a_1} x_2^{a_2} \ldots x_n^{a_n}$ $(a_i \geq 0)$ to be $a_1 + a_2 + \cdots + a_n$. Recall that a polynomial is *homogeneous (of weight m)* if every monomial term has the same weight (m).

37. **(a)** Show that every polynomial $f(x_1, \ldots, x_n)$ can be written as a sum of homogeneous polynomials. Show that if $f(x_1, \ldots, x_n)$ is symmetric then each of these homogeneous polynomials is also symmetric.

(b) Show that the monomial $Bs_1^{a_1} s_2^{a_2} \ldots s_n^{a_n}$ in the elementary symmetric functions is a homogenous polynomial in $x_1, x_2, \ldots, x_n$ of weight $a_1 + 2a_2 + \cdots + na_n$.

In writing $f(x_1, \ldots, x_n)$ as a polynomial in the symmetric functions it therefore suffices to assume that $f(x_1, \ldots, x_n)$ is homogeneous.

We can *lexicographically order* the monomial terms similarly to the ordering in a dictionary (hence the name) by declaring the term whose exponents are $(a_1, a_2, \ldots, a_n)$ to come before the term with exponents $(b_1, b_2, \ldots, b_n)$ if the initial terms of the two n-tuples are equal and the first term where they differ has $a_i > b_i$. If $f(x_1, \ldots, x_n)$ contains the monomial term $Ax_1^{a_1} x_2^{a_2} \ldots x_n^{a_n}$ then since $f(x_1, \ldots, x_n)$ is symmetric it also contains all the permuted monomial terms. Among these terms choose the lexicographically largest term, which therefore satisfies $a_1 \geq a_2 \geq \cdots \geq a_n \geq 0$.

38. **(a)** Show that the monomial $As_1^{a_1 - a_2} s_2^{a_2 - a_3} \ldots s_n^{a_n}$ in the elementary symmetric functions has the same lexicographic initial term.

(b) Show that subtracting $As_1^{a_1 - a_2} s_2^{a_2 - a_3} \ldots s_n^{a_n}$ from $f(x)$ yields either 0 or a symmetric polynomial of the same weight whose terms are lexicographically smaller than the terms in $f(x_1, \ldots, x_n)$.

(c) Show that the iteration of this procedure (lexicographic ordering, choosing the lexicographically largest term, subtracting the associated monomial in the elementary symmetric functions) terminates, expressing $f(x_1, \ldots, x_n)$ as a polynomial in the elementary symmetric functions.

39. Use the algorithm described in the previous exercise to prove that a polynomial $f(x_1, \ldots, x_n)$ that is symmetric in $x_1, \ldots, x_n$ can be expressed *uniquely* as a polynomial in the elementary symmetric functions.

40. Use the procedure in Exercise 38 to express each of the following symmetric functions as a polynomial in the elementary symmetric functions:
    (a) $(x_1 - x_2)^2$
    (b) $x_1^2 + x_2^2 + x_3^2$
    (c) $x_1^2x_2^2 + x_1^2x_3^2 + x_2^2x_3^2$.

41. Use the procedure in Exercise 38 to express $\sum_{i \neq j} x_i^2 x_j$ as a polynomial in the elementary symmetric functions.

We now know that a symmetric polynomial $f(x_1, \ldots, x_n)$ can be written uniquely as a polynomial in the elementary symmetric functions. Using this existence and uniqueness we can describe an alternate and computationally useful method for determining the coefficients of the elementary symmetric functions in this polynomial. As in Exercise 37 we may assume that $f(x_1, \ldots, x_n)$ is homogeneous of weight M. Let N be the maximum degree of any of the variables $x_1, \ldots, x_n$ in $f(x_1, \ldots, x_n)$.

(a) Determine all of the possible monomial terms $A_i s_1^{a_1} s_2^{a_2} \ldots s_n^{a_n}$ appearing in $f(x_1, \ldots, x_n)$ from the constraints

$$a_1 + 2a_2 + \ldots + na_n = M$$
$$a_1 + a_2 + \ldots + a_n \leq N\,.$$

(b) Since $f(x_1, \ldots, x_n) = \sum A_i s_1^{a_1} s_2^{a_2} \ldots s_n^{a_n}$ is a polynomial *identity*, it is valid for any substitution of values for $x_1, \ldots, x_n$. Each substitution into this equation gives a linear relation on the coefficients A_i and so a sufficient number of substitutions will determine the A_i.

42. Show that the function $(x_1 + x_2 - x_3 - x_4)(x_1 + x_3 - x_2 - x_4)(x_1 + x_4 - x_2 - x_3)$ is symmetric in x_1, x_2, x_3, x_4 and use the preceding procedure to prove it can be expressed as a polynomial in the elementary symmetric functions as $s_1^3 - 4s_1s_2 + 8s_3$.

43. Express each of the following in terms of the elementary symmetric functions:
    (a) $\sum_{i \neq j} x_i^2 x_j$ (b) $\sum_{i,j,k \text{ distinct}} x_i^2 x_j x_k$ (c) $\sum_{i,j,k \text{ distinct}} x_i^2 x_j^2 x_k^2$.

44. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ be the roots of a quartic polynomial $f(x)$ over $\mathbb{Q}$. Show that the quantities $\alpha_1\alpha_2 + \alpha_3\alpha_4$, $\alpha_1\alpha_3 + \alpha_2\alpha_4$, and $\alpha_1\alpha_4 + \alpha_2\alpha_3$ are permuted by the Galois group of $f(x)$. Conclude that these elements are the roots of a cubic polynomial with coefficients in $\mathbb{Q}$ (also sometimes referred to as the *resolvent cubic* of $f(x)$).

45. If $f(x) = x^3 + px + q \in \mathbb{Z}[x]$ is irreducible, prove that its discriminant $D = -4p^3 - 27q^2$ is an integer not equal to $0, \pm 1$.

46. Prove that every finite group occurs as the Galois group of some field extension $F(x_1, x_2, \ldots, x_n)/E$.

47. Let F be a field of characteristic 0 in which every cubic polynomial has a root. Let $f(x)$ be an irreducible quartic polynomial over F whose discriminant is a square in F. Determine the Galois group of $f(x)$.

48. This exercise determines the splitting field K for the polynomial $f(x) = x^6 - 2x^3 - 2$ over $\mathbb{Q}$ (cf. also Exercise 2 of Section 8).
    (a) Prove that $f(x)$ is irreducible over $\mathbb{Q}$ with roots the three cube roots of $1 \pm \sqrt{3}$.
    (b) Prove that K contains the field $\mathbb{Q}(\sqrt{-3})$ of 3rd roots of unity and contains $\mathbb{Q}(\sqrt{3})$, hence contains the biquadratic field $F = \mathbb{Q}(i, \sqrt{3})$. Take the product of two of the

roots in (a) to prove that K contains $\sqrt[3]{2}$ and conclude that K is an extension of the field $L = \mathbb{Q}(\sqrt[3]{2}, i, \sqrt{3})$.

(c) Prove that $[L : \mathbb{Q}] = 12$ and that K is obtained from L by adjoining the cube root of an element in L, so that $[K : \mathbb{Q}] = 12$ or 36.

(d) Prove that if $[K : \mathbb{Q}] = 12$ then $K = \mathbb{Q}(\sqrt[3]{2}, i, \sqrt{3})$ and that $\mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to the direct product of the cyclic group of order 2 and S_3. Prove that if $[K : \mathbb{Q}] = 12$ then there is a unique real cubic subfield in K, namely $\mathbb{Q}(\sqrt[3]{2})$.

(e) Take the quotient of the two real roots in (a) to show that $\sqrt[3]{2+\sqrt{3}}$ and $\sqrt[3]{2-\sqrt{3}}$ (real roots) are both elements of K. Show that $\alpha = \sqrt[3]{2+\sqrt{3}} + \sqrt[3]{2-\sqrt{3}}$ is a real root of the irreducible cubic equation $x^3 - 3x - 4$ whose discriminant is $-2^2 3^4$. Conclude that the Galois closure of $\mathbb{Q}(\alpha)$ contains $\mathbb{Q}(i)$ so in particular $\mathbb{Q}(\alpha) \neq \mathbb{Q}(\sqrt[3]{2})$.

(f) Conclude from (e) that $G = \mathrm{Gal}(K/\mathbb{Q})$ is of order 36. Determine all the elements of G explicitly and in particular show that G is isomorphic to $S_3 \times S_3$.

49. Prove that the Galois group over $\mathbb{Q}$ of $x^6 - 4x^3 + 1$ is isomorphic to the dihedral group of order 12. [Observe that the two real roots are inverses of each other.]

14.7 SOLVABLE AND RADICAL EXTENSIONS: INSOLVABILITY OF THE QUINTIC

We now investigate the question of solving for the roots of a polynomial by *radicals,* that is, in terms of the algebraic operations of addition, subtraction, multiplication, division and the extraction of nth roots. The quadratic formula for the roots of a polynomial of degree 2 is familiar from elementary algebra and we shall derive below similar formulas for the roots of cubic and quartic polynomials. For polynomials of degree ≥ 5, however, we shall see that such formulas are not possible – this is Abel's Theorem on the insolvability of the general quintic. The reason for this is quite simple: we shall see that a polynomial is solvable by radicals if and only if its Galois group is a solvable group (which explains the terminology) and for $n \geq 5$ the group S_n is not solvable.

We first discuss *simple radical extensions,* namely extensions obtained by adjoining to a field F the nth root of an element a in F. Since all the roots of the polynomial $x^n - a$ for $a \in F$ differ by the nth roots of unity, adjoining one such root will give a Galois extension if and only if this field contains the nth roots of unity. Simple radical extensions are best behaved when the base field F already contains the appropriate roots of unity. The symbol $\sqrt[n]{a}$ for $a \in F$ will be used to denote any root of the polynomial $x^n - a \in F[x]$.

Definition. The extension K/F is said to be *cyclic* if it is Galois with a cyclic Galois group.

Proposition 36. Let F be a field of characteristic not dividing n which contains the nth roots of unity. Then the extension $F(\sqrt[n]{a})$ for $a \in F$ is cyclic over F of degree dividing n.

Proof: The extension $K = F(\sqrt[n]{a})$ is Galois over F if F contains the nth roots of unity since it is the splitting field for $x^n - a$. For any $\sigma \in \mathrm{Gal}(K/F)$, $\sigma(\sqrt[n]{a})$ is

another root of this polynomial, hence $\sigma(\sqrt[n]{a}) = \zeta_\sigma \sqrt[n]{a}$ for some n^{th} root of unity ζ_σ. This gives a map

$$\begin{aligned} \mathrm{Gal}(K/F) &\to \mu_n \\ \sigma &\mapsto \zeta_\sigma \end{aligned}$$

where μ_n denotes the group of nth roots of unity. Since F contains μ_n, every n^{th} root of unity is fixed by every element of $\mathrm{Gal}(K/F)$. Hence

$$\begin{aligned} \sigma\tau(\sqrt[n]{a}) &= \sigma(\zeta_\tau \sqrt[n]{a}) \\ &= \zeta_\tau \sigma(\sqrt[n]{a}) \\ &= \zeta_\tau \zeta_\sigma \sqrt[n]{a} = \zeta_\sigma \zeta_\tau \sqrt[n]{a} \end{aligned}$$

which shows that $\zeta_{\sigma\tau} = \zeta_\sigma \zeta_\tau$, so the map above is a homomorphism. The kernel consists precisely of the automorphisms which fix $\sqrt[n]{a}$, namely the identity. This gives an injection of $\mathrm{Gal}(K/F)$ into the cyclic group μ_n of order n, which proves the proposition.

Let now K be any cyclic extension of degree n over a field F of characteristic not dividing n which contains the nth roots of unity. Let σ be a generator for the cyclic group $\mathrm{Gal}(K/F)$.

Definition. For $\alpha \in K$ and any n^{th} root of unity ζ, define the *Lagrange resolvent* $(\alpha, \zeta) \in K$ by

$$(\alpha, \zeta) = \alpha + \zeta\sigma(\alpha) + \zeta^2\sigma^2(\alpha) + \cdots + \zeta^{n-1}\sigma^{n-1}(\alpha).$$

If we apply the automorphism σ to (α, ζ) we obtain

$$\sigma(\alpha, \zeta) = \sigma\alpha + \zeta\sigma^2(\alpha) + \zeta^2\sigma^3(\alpha) + \cdots + \zeta^{n-1}\sigma^n(\alpha)$$

since ζ is an element of the base field F so is fixed by σ. We have $\zeta^n = 1$ in μ_n and $\sigma^n = 1$ in $\mathrm{Gal}(K/F)$ so this can be written

$$\begin{aligned} \sigma(\alpha, \zeta) &= \sigma\alpha + \zeta\sigma^2(\alpha) + \zeta^2\sigma^3(\alpha) + \cdots + \zeta^{-1}\alpha \\ &= \zeta^{-1}(\alpha + \zeta\sigma(\alpha) + \zeta^2\sigma^2(\alpha) + \cdots + \zeta^{n-1}\sigma^{n-1}(\alpha)) \\ &= \zeta^{-1}(\alpha, \zeta). \end{aligned} \tag{14.19}$$

It follows that

$$\sigma(\alpha, \zeta)^n = (\zeta^{-1})^n(\alpha, \zeta)^n = (\alpha, \zeta)^n$$

so that $(\alpha, \zeta)^n$ is fixed by $\mathrm{Gal}(K/F)$, hence is an element of F for any $\alpha \in K$.

Let ζ be a primitive n^{th} root of unity. By the linear independence of the automorphisms $1, \sigma, \ldots, \sigma^{n-1}$ (Theorem 7), there is an element $\alpha \in K$ with $(\alpha, \zeta) \neq 0$. Iterating (19) we have

$$\sigma^i(\alpha, \zeta) = \zeta^{-i}(\alpha, \zeta), \qquad i = 0, 1, \ldots,$$

and it follows that σ^i does not fix (α, ζ) for any $i < n$. Hence this element cannot lie in any proper subfield of K, so $K = F((\alpha, \zeta))$. Since we proved $(\alpha, \zeta)^n = a \in F$ above, we have $F(\sqrt[n]{a}) = F((\alpha, \zeta)) = K$. This proves the converse of Proposition 36:

Proposition 37. Any cyclic extension of degree n over a field F of characteristic not dividing n which contains the nth roots of unity is of the form $F(\sqrt[n]{a})$ for some $a \in F$.

Remark: The two propositions above form a part of what is referred to as *Kummer theory.* A group G is said to have *exponent* n if $g^n = 1$ for every $g \in G$. Let F be a field of characteristic not dividing n which contains the nth roots of unity. If we take elements $a_1, ..., a_k \in F^\times$ then as in Proposition 36 we can see that the extension

$$F(\sqrt[n]{a_1}, \sqrt[n]{a_2}, ..., \sqrt[n]{a_k}) \tag{14.20}$$

is an abelian extension of F whose Galois group is of exponent n. Conversely, any abelian extension of exponent n is of this form.

Denote by $(F^\times)^n$ the subgroup of the multiplicative group $F^\times$ consisting of the nth powers of nonzero elements of F. The quotient group $F^\times/(F^\times)^n$ is an abelian group of exponent n. The Galois group of the extension in (20) is isomorphic to the group generated in $F^\times/(F^\times)^n$ by the elements $a_1, ..., a_k$ and two extensions as in (20) are equal if and only if their associated groups in $F^\times/(F^\times)^n$ are equal.

Hence the (finitely generated) subgroups of $F^\times/(F^\times)^n$ classify the abelian extensions of exponent n over fields containing the nth roots of unity (and characteristic not dividing n). Such extensions are called *Kummer extensions*.

These results generalize the case $k = 1$ above and can be proved in a similar way.

For simplicity we now consider the situation of a base field F of characteristic 0. As in the previous propositions the results are valid over fields whose characteristics do not divide any of the orders of the roots that will be taken.

Definition

(1) An element α which is algebraic over F can be *expressed by radicals* or *solved for in terms of radicals* if α is an element of a field K which can be obtained by a succession of simple radical extensions:

$$F = K_0 \subset K_1 \subset ... \subset K_i \subset K_{i+1} \subset ... \subset K_s = K \tag{14.21}$$

where $K_{i+1} = K_i(\sqrt[n_i]{a_i})$ for some $a_i \in K_i$, $i = 0, 1, ..., s-1$. Here $\sqrt[n_i]{a_i}$ denotes some root of the polynomial $x^{n_i} - a_i$. Such a field K will be called a *root extension* of F.

(2) A polynomial $f(x) \in F[x]$ can be *solved by radicals* if all its roots can be solved for in terms of radicals.

This gives a precise meaning to the intuitive notion that α is obtained by successive algebraic operations (addition, subtraction, multiplication and division) and successive root extractions. For example, the element

$$-1 + \sqrt{17} + \sqrt{2(17 - \sqrt{17})} + 2\sqrt{17 + 3\sqrt{17} - \sqrt{2(17 - \sqrt{17})} - 2\sqrt{2(17 + \sqrt{17})}}.$$

encountered at the end of Section 5 (used to construct the regular 17-gon) is expressed by radicals and is contained in the field K_4, where

$$\begin{aligned}
K_0 &= \mathbb{Q} \\
K_1 &= K_0(\sqrt{a_0}) & a_0 &= 17 \\
K_2 &= K_1(\sqrt{a_1}) & a_1 &= 2(17 - \sqrt{17}) \\
K_3 &= K_2(\sqrt{a_2}) & a_2 &= 2(17 + \sqrt{17}) \\
K_4 &= K_3(\sqrt{a_3}) & a_3 &= 17 + 3\sqrt{17} - \sqrt{2(17 - \sqrt{17})} - 2\sqrt{2(17 + \sqrt{17})}.
\end{aligned}$$

Each of these extensions is a radical extension. The fact that no roots other than square roots are required reflects the fact that the regular 17-gon is constructible by straightedge and compass.

In considering radical extensions one may always adjoin roots of unity, since by definition the roots of unity are radicals. This is useful because then cyclic extensions become radical extensions and conversely. In particular we have

Lemma 38. If α is contained in a root extension K as in (21) above, then α is contained in a root extension which is Galois over F and where each extension K_{i+1}/K_i is cyclic.

Proof: Let L be the Galois closure of K over F. For any $\sigma \in \mathrm{Gal}(L/F)$ we have the chain of subfields

$$F = \sigma K_0 \subset \sigma K_1 \subset \ldots \subset \sigma K_i \subset \sigma K_{i+1} \subset \ldots \subset \sigma K_s = \sigma K$$

where $\sigma K_{i+1}/\sigma K_i$ is again a simple radical extension (since it is generated by the element $\sigma(\sqrt[n_i]{a_i})$, which is a root of the equation $x^{n_i} - \sigma(a_i)$ over $\sigma(K_i)$). It is easy to see that the composite of two root extensions is again a root extension (if K' is another root extension with subfields K'_i, first take the composite of K'_1 with the fields $K_0, K_1, \ldots, K_s$, then the composite of these fields with K'_2, etc. so that each individual extension in this process is a simple radical extension). It follows that the composite of all the conjugate fields $\sigma(K)$ for $\sigma \in \mathrm{Gal}(L/F)$ is again a root extension. Since this field is precisely L, we see that α is contained in a Galois root extension.

We now adjoin to F the n_i-th roots of unity for all the roots $\sqrt[n_i]{a_i}$ of the simple radical extensions in the Galois root extension K/F, obtaining the field F', say, and then form the composite of F' with the root extension:

$$F \subseteq F' = F'K_0 \subseteq F'K_1 \subseteq \cdots \subseteq F'K_i \subseteq F'K_{i+1} \subseteq \cdots \subseteq F'K_s = F'K.$$

The field $F'K$ is a Galois extension of F since it is the composite of two Galois extensions. The extension from F to $F' = F'K_0$ can be given as a chain of subfields with each individual extension cyclic (this is true for any abelian extension). Each extension $F'K_{i+1}/F'K_i$ is a simple radical extension and since we now have the appropriate roots of unity in the base fields, each of these individual extensions from F' to $F'K$ is a cyclic extension by Proposition 36. Hence $F'K/F$ is a root extension which is Galois over F with cyclic intermediate extensions, completing the proof.

Recall from Section 3.4 (cf. also Section 6.1) that a finite group G is *solvable* if there exists a chain of subgroups

$$1 = G_s \leq G_{s-1} \leq \ldots \leq G_{i+1} \leq G_i \leq \ldots \leq G_0 = G \tag{14.22}$$

with G_i/G_{i+1} cyclic, $i = 0, 1, \ldots, s-1$. We have proved that subgroups and quotient groups of solvable groups are solvable and that if $H \leq G$ and G/H are both solvable, then G is solvable.

We now prove Galois' fundamental connection between solving for the roots of polynomials in terms of radicals and the Galois group of the polynomial. We continue to work over a field F of characteristic 0, but it is easy to see that the proof is valid over any field of characteristic not dividing the order of the Galois group or the orders of the radicals involved.

Theorem 39. The polynomial $f(x)$ can be solved by radicals if and only if its Galois group is a solvable group.

Proof: Suppose first that $f(x)$ can be solved by radicals. Then each root of $f(x)$ is contained in an extension as in the lemma. The composite L of such extensions is again of the same type by Proposition 21. Let G_i be the subgroups corresponding to the subfields K_i, $i = 0, 1, \ldots, s-1$. Since

$$\mathrm{Gal}(K_{i+1}/K_i) = G_i/G_{i+1} \qquad i = 0, 1, \ldots, s-1$$

it follows that the Galois group $G = \mathrm{Gal}(L/F)$ is a solvable group. The field L contains the splitting field of $f(x)$ so the Galois group of $f(x)$ is a quotient group of the solvable group G, hence is solvable.

Suppose now that the Galois group G of $f(x)$ is a solvable group and let K be the splitting field for $f(x)$. Taking the fixed fields of the subgroups in a chain (22) for G gives a chain

$$F = K_0 \subset K_1 \subset \ldots \subset K_i \subset K_{i+1} \subset \ldots \subset K_s = K$$

where K_{i+1}/K_i, $i = 0, 1, \ldots, s-1$ is a cyclic extension of degree n_i. Let F' be the cyclotomic field over F of all roots of unity of order n_i, $i = 0, 1, \ldots, s-1$ and form the composite fields $K_i' = F'K_i$. We obtain a sequence of extensions

$$F \subseteq F' = F'K_0 \subseteq F'K_1 \subseteq \cdots \subseteq F'K_i \subseteq F'K_{i+1} \subseteq \cdots \subseteq F'K_s = F'K\,.$$

The extension $F'K_{i+1}/F'K_i$ is cyclic of degree dividing n_i, $i = 0, 1, \ldots, s-1$ (by Proposition 19). Since we now have the appropriate roots of unity in the base fields, each of these cyclic extensions is a simple radical extension by Proposition 37. Each of the roots of $f(x)$ is therefore contained in the root extension $F'K$ so that $f(x)$ can be solved by radicals.

Corollary 40. The general equation of degree n cannot be solved by radicals for $n \geq 5$.

Proof: For $n \geq 5$ the group S_n is not solvable as we showed in Chapter 4. The corollary follows immediately from Theorem 32.

This corollary shows that there is no formula involving radicals analogous to the quadratic formula for polynomials of degree 2 for the roots of a polynomial of degree 5. To give an example of a *specific* polynomial over $\mathbb{Q}$ of degree 5 whose roots cannot be expressed in terms of radicals we must demonstrate a polynomial of degree 5 with rational coefficients having S_5 (or A_5, which is also not solvable) as Galois group (cf. also Exercise 21, which gives a criterion for the solvability of a quintic).

Example

Consider the polynomial $f(x) = x^5 - 6x + 3 \in \mathbb{Q}[x]$. This polynomial is irreducible since it is Eisenstein at 3. The splitting field K for this polynomial therefore has degree divisible by 5, since adjoining one root of $f(x)$ to $\mathbb{Q}$ generates an extension of degree 5. The Galois group G is therefore a subgroup of S_5 of order divisible by 5 so contains an element of order 5. The only elements in S_5 of order 5 are 5-cycles, so G contains a 5-cycle.

Since $f(-2) = -17$, $f(0) = 3$, $f(1) = -2$, and $f(2) = 23$ we see that $f(x)$ has a real root in each of the intervals $(-2, 0)$, $(0, 1)$ and $(1, 2)$. By the Mean Value Theorem, if there were 4 real roots then the derivative $f'(x) = 5x^4 - 6$ would have at least 3 real zeros, which it does not. Hence these are the only real roots. (This also follows easily by Descarte's rule of signs.) By the Fundamental Theorem of Algebra $f(x)$ has 5 roots in $\mathbb{C}$. Hence $f(x)$ has two complex roots which are not real. Let τ denote the automorphism of complex conjugation in $\mathbb{C}$. Since the coefficients of $f(x)$ are real, the two complex roots must be interchanged by τ (since they are not fixed, not being real). Hence the restriction of complex conjugation to K fixes three of the roots of $f(x)$ and interchanges the other two. As an element of G, $\tau|_K$ is therefore a transposition.

It is now a simple exercise to show that any 5-cycle together with any transposition generate all of S_5. It follows that $G = S_5$, so the roots of $x^5 - 6x + 3$ cannot be expressed by radicals.

As indicated in this example, a great deal of information regarding the Galois group can be obtained by understanding the *cycle types* of the automorphisms in G considered as a subgroup of S_n. In practice this is the most efficient way of determining the Galois groups of polynomials of degrees ≥ 5 (becoming more difficult the larger the degree, of course, if only because the possible subgroups of S_n are vastly more numerous). We describe this procedure in the next section.

By Theorem 39, any polynomial of degree $n \leq 4$ can be solved by radicals, since S_n is a solvable group for these n. For $n = 2$ this is just the familiar quadratic formula. For $n = 3$ the formula is known as *Cardano's formula* (named for Geronimo Cardano (1501–1576)) and the formula for $n = 4$ can be reduced to this one. The formulas are valid over any field F of characteristic $\neq 2, 3$, which are the characteristics dividing the orders of the radicals necessary and the orders of the possible Galois groups (which are subgroups of S_3 and S_4). For simplicity we shall derive the formulas over $\mathbb{Q}$.

Solution of Cubic Equations by Radicals: Cardano's Formulas

From the proof of Theorem 39 and the fact that a composition series for S_3 as in equation (22) is given by $1 \leq A_3 \leq S_3$ we should expect that the solution of the cubic

$$f(x) = x^3 + ax^2 + bx + c$$

(or equivalently, under the substitution $x = y - a/3$,

$$g(y) = y^3 + py + q,$$

where

$$p = \frac{1}{3}(3b - a^2) \qquad q = \frac{1}{27}(2a^3 - 9ab + 27c) \qquad)$$

to involve adjoining the 3rd roots of unity and the formation of Lagrange resolvents involving these roots of unity.

Let ρ denote a primitive 3rd root of unity, so that $\rho^2 + \rho + 1 = 0$. Let the roots of $g(y)$ be α, β, and γ, so that

$$\alpha + \beta + \gamma = 0$$

(one of the reasons for changing from $f(x)$ to $g(x)$). Over the field $\mathbb{Q}(\sqrt{D})$ where D is the discriminant (computed in the last section) the Galois group of $g(y)$ is A_3, i.e. a cyclic group of order 3. If we adjoin ρ then this extension is a radical extension of degree 3, with generator given by a Lagrange Resolvent, as in the proof of Proposition 37. Consider therefore the elements

$$\begin{aligned}(\alpha, 1) &= \alpha + \beta + \gamma = 0\\ \theta_1 = (\alpha, \rho) &= \alpha + \rho\beta + \rho^2\gamma\\ \theta_2 = (\alpha, \rho^2) &= \alpha + \rho^2\beta + \rho\gamma.\end{aligned}$$

Note that the sum of these resolvents is

$$\theta_1 + \theta_2 = 3\alpha \tag{14.23}$$

since $1 + \rho + \rho^2 = 0$. Similarly

$$\begin{aligned}\rho^2\theta_1 + \rho\theta_2 &= 3\beta\\ \rho\theta_1 + \rho^2\theta_2 &= 3\gamma.\end{aligned} \tag{14.23$'$}$$

We also showed in general before Proposition 37 that the cube of these resolvents must lie in $\mathbb{Q}(\sqrt{D}, \rho)$. Expanding θ_1^3 we obtain

$$\begin{aligned}\alpha^3 + \beta^3 + \gamma^3 &+ 3\rho(\alpha^2\beta + \beta^2\gamma + \alpha\gamma^2)\\ &+ 3\rho^2(\alpha\beta^2 + \beta\gamma^2 + \alpha^2\gamma) + 6\alpha\beta\gamma.\end{aligned} \tag{14.24}$$

We have

$$\begin{aligned}\sqrt{D} &= (\alpha - \beta)(\alpha - \gamma)(\beta - \gamma)\\ &= (\alpha^2\beta + \beta^2\gamma + \alpha\gamma^2) - (\alpha\beta^2 + \beta\gamma^2 + \alpha^2\gamma).\end{aligned}$$

Using this equation we see that (24) can be written

$$\alpha^3 + \beta^3 + \gamma^3 + 3\rho[\frac{1}{2}(S + \sqrt{D})] + 3\rho^2[\frac{1}{2}(S - \sqrt{D})] + 6\alpha\beta\gamma \tag{14.24$'$}$$

where for simplicity we have denoted by S the expression

$$(\alpha^2\beta + \beta^2\gamma + \alpha\gamma^2) + (\alpha\beta^2 + \beta\gamma^2 + \alpha^2\gamma).$$

Since S is symmetric in the roots, each of the expressions in (24′) is a symmetric polynomial in α, β and γ, hence is a polynomial in the elementary symmetric functions $s_1 = 0$, $s_2 = p$, and $s_3 = -q$. After a short calculation one finds

$$\alpha^3 + \beta^3 + \gamma^3 = -3q \qquad S = 3q$$

so that from (24′) we find ($\rho + \rho^2 = -1$ and $\rho - \rho^2 = \sqrt{-3}$)

$$\begin{aligned}\theta_1^3 &= -3q + \frac{3}{2}\rho(3q + \sqrt{D}) + \frac{3}{2}\rho^2(3q - \sqrt{D}) - 6q \\ &= \frac{-27}{2}q + \frac{3}{2}\sqrt{-3D}.\end{aligned} \tag{14.25}$$

Similarly, we find

$$\theta_2^3 = \frac{-27}{2}q - \frac{3}{2}\sqrt{-3D}. \tag{14.25′}$$

Equations (25) and (23) essentially give the solutions of our cubic. One small point remains, however, namely the issue of extracting the cube roots of the expressions in (25) to obtain θ_1 and θ_2. There are 3 possible cube roots, which might suggest a total of 9 expressions in (23). This is not the case since θ_1 and θ_2 are not independent (adjoining one of them already gives the Galois extension containing all of the roots). A computation like the one above (but easier) shows that

$$\theta_1\theta_2 = -3p \tag{14.26}$$

showing that the choice of cube root for θ_1 determines θ_2. Using $D = -4p^3 - 27q^2$, we obtain Cardano's explicit formulas: Let

$$\begin{aligned}A &= \sqrt[3]{\frac{-27}{2}q + \frac{3}{2}\sqrt{-3D}} \\ B &= \sqrt[3]{\frac{-27}{2}q - \frac{3}{2}\sqrt{-3D}}\end{aligned}$$

where the cube roots are chosen so that $AB = -3p$. Then the roots of the equation

$$y^3 + py + q = 0$$

are

$$\alpha = \frac{A + B}{3} \qquad \beta = \frac{\rho^2 A + \rho B}{3} \qquad \gamma = \frac{\rho A + \rho^2 B}{3} \tag{14.27}$$

where $\rho = -\frac{1}{2} + \frac{1}{2}\sqrt{-3}$.

Examples

(1) Consider the cubic equation $x^3 - x + 1 = 0$. The discriminant of this cubic is

$$D = -4(-1)^3 - 27(1)^2 = -23$$

which is not the square of a rational number, so the Galois group for this polynomial is S_3. Substituting into the formulas above we have

$$A = \sqrt[3]{\frac{-27}{2} + \frac{3}{2}\sqrt{69}}$$

$$B = \sqrt[3]{\frac{-27}{2} - \frac{3}{2}\sqrt{69}}$$

where we choose A to be the real cube root and then from $AB = 3$ we see that B is also real. The roots of the cubic are given by (27) and we see that there is one real root and two (conjugate) complex roots (which we could have determined without solving for the roots, of course).

(2) Consider the equation $x^3 + x^2 - 2x - 1 = 0$. Letting $x = s - 1/3$ the equation becomes $s^3 - \frac{7}{3}s - \frac{7}{27} = 0$. Multiplying through by 27 to clear denominators and letting $y = 3s$ we see that y satisfies the cubic equation

$$y^3 - 21y - 7 = 0.$$

The discriminant D for this cubic is

$$D = -4(-21)^3 - 27(-7)^2 = 3^6 7^2$$

which shows that the Galois group for this (Eisenstein at 7) cubic is A_3. Substituting into the formulas above we have

$$A = 3\sqrt[3]{\frac{7}{2} + \frac{21}{2}\sqrt{-3}}$$

$$B = 3\sqrt[3]{\frac{7}{2} - \frac{21}{2}\sqrt{-3}}$$

and the roots of our cubics can be expressed in terms of A and B using the formulas above. This cubic arises from trying to express a primitive 7th root of unity ζ_7 in terms of radicals similar to the explicit formulas for the other roots of unity of small order (cf. the exercises).

In this case we have $g(-5) = -27$, $g(-1) = 13$, $g(0) = -7$ and $g(5) = 13$, so that this cubic has 3 *real* roots. The expressions above for these roots are sums of the conjugates of *complex* numbers. We shall see below that this is necessary, namely that it is impossible to solve for these real roots using only radicals involving real numbers.

A cubic with rational coefficients has either one real root and two complex conjugate imaginary roots or has three real roots. These two cases can be distinguished by the sign of the discriminant:

Suppose in the first case that the roots are a and $b \pm ic$ where a, b, and c are real and $c \neq 0$. Then

$$\begin{aligned}\sqrt{D} &= [a - (b + ic)][a - (b - ic)][(b + ic) - (b - ic)] \\ &= 2ic[(a - b)^2 + c^2]\end{aligned}$$

is purely imaginary, so that the discriminant D is negative. Then in the formulas for A and B above we may choose both to be real. The first root in (27) is then real and the second two are complex conjugates.

If all three roots are real, then clearly $\sqrt{D}$ is real, so $D \geq 0$ is a nonnegative real number. If $D = 0$ then the cubic has repeated roots. For $D > 0$ (sometimes called the *Casus irreducibilis*), the formulas for the roots involve radicals of nonreal numbers, as in Example 2. We now show that for irreducible cubics this is necessary. The exercises outline the proof of the following generalization: if all the roots of the irreducible polynomial $f(x) \in \mathbb{Q}[x]$ are real and if one of these roots can be expressed by *real* radicals, then the degree of $f(x)$ is a power of 2, the Galois group of $f(x)$ is a 2-group, and the roots of $f(x)$ can be constructed by straightedge and compass.

Suppose that the irreducible cubic $f(x)$ has three real roots and that it were possible to express one of these roots by radicals involving only real numbers. Then the splitting field for the cubic would be contained in a root extension

$$\mathbb{Q} = K_0 \subset K_1 = \mathbb{Q}(\sqrt{D}) \subset \ldots \subset K_i \subset K_{i+1} \subset \ldots \subset K_s = K$$

where each field K_i, $i = 0, 1, \ldots, s$, is contained in the real numbers $\mathbb{R}$ and $s \geq 2$ since the quadratic extension $\mathbb{Q}(\sqrt{D})$ cannot contain the root of an irreducible cubic. We have begun this root extension with $\mathbb{Q}(\sqrt{D})$ because over this field the Galois group of the polynomial is cyclic of degree 3.

Note that for any field F the extension $F(\sqrt[mn]{a})$ of F can be obtained by two smaller simple radical extensions: let

$$F_1 = F(\sqrt[n]{a})$$

and let $b = \sqrt[n]{a} \in F_1$, so that

$$F(\sqrt[mn]{a}) = F_1(\sqrt[m]{b}).$$

We may therefore always assume our radical extensions are of the form $F(\sqrt[p]{a})$ where p is a prime.

Suppose now that F is a subfield of the real numbers $\mathbb{R}$ and let a be an element of F. Let p be a prime and let $\alpha = \sqrt[p]{a}$ denote a real p^{th} root of a. Then $[F(\sqrt[p]{a}) : F]$ must be either 1 or p, as follows. The conjugates of α over F all differ from α by a pth root of unity. It follows that the constant term of the minimal polynomial of α over F is $\alpha^d\zeta$ where $d = [F(\sqrt[p]{a}) : F]$ is the degree of the minimal polynomial and ζ is some p^{th} root of unity. Since α is real and $\alpha^d\zeta \in F$ is real, it follows that $\zeta = \pm 1$, so that $\alpha^d \in F$. Then, if $d \neq p$, $\alpha^d \in F$ and $\alpha^p = a \in F$ implies $\alpha \in F$, so $d = 1$.

Hence we may assume for the radical extensions above that $[K_{i+1} : K_i]$ is a prime p_i and $K_{i+1} = K_i(\sqrt[p_i]{a_i})$ for some $a_i \in K_i$, $i = 0, 1, \ldots, s-1$. In other words, the original tower of real radical extensions can be refined to a tower where each of the successive radical extensions has prime degree.

If any field containing $\sqrt{D}$ contains one of the roots of $f(x)$ then it contains the splitting field for $f(x)$, hence contains all the roots of the cubic. We suppose s chosen so that K_{s-1} does not contain any of the roots of the cubic.

Consider the extension K_s/K_{s-1}. The field K_s contains all the roots of the cubic $f(x)$ and the field K_{s-1} contains none of these roots. It follows that $f(x)$ is irreducible

over K_{s-1}, so $[K_s : K_{s-1}]$ is divisible by 3. Since we have reduced to the case where this extension degree is a prime, it follows that the extension degree is precisely 3 and that the extension K_s/K_{s-1} is Galois (being the splitting field of $f(x)$ over K_{s-1}). Since also $K_s = K_{s-1}(\sqrt[3]{a})$ for some $a \in K_{s-1}$, the Galois extension K_s must also contain the other cube roots of a. This implies that K_s contains ρ, a primitive 3rd root of unity. This contradicts the assumption that K_s is a subfield of $\mathbb{R}$ and shows that it is impossible to express the roots of this cubic in terms of real radicals only.

Solution of Quartic Equations by Radicals

Consider now the case of a quartic polynomial

$$f(x) = x^4 + ax^3 + bx^2 + cx + d$$

which under the substitution $x = y - a/4$ becomes the quartic

$$g(y) = y^4 + py^2 + qy + r$$

with

$$\begin{aligned} p &= \frac{1}{8}(-3a^2 + 8b) \\ q &= \frac{1}{8}(a^3 - 4ab + 8c) \\ r &= \frac{1}{256}(-3a^4 + 16a^2b - 64ac + 256d). \end{aligned}$$

Let the roots of $g(y)$ be α_1, α_2, α_3, and α_4. The resolvent cubic introduced in the previous section for this quartic is

$$h(x) = x^3 - 2px^2 + (p^2 - 4r)x + q^2$$

and has roots

$$\begin{aligned} \theta_1 &= (\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4) \\ \theta_2 &= (\alpha_1 + \alpha_3)(\alpha_2 + \alpha_4) \\ \theta_3 &= (\alpha_1 + \alpha_4)(\alpha_2 + \alpha_3). \end{aligned}$$

The Galois group of the splitting field for $f(x)$ (or $g(y)$) over the splitting field of the resolvent cubic $h(x)$ is the Klein 4-group. Such extensions are biquadratic, which means that it is possible to solve for the roots α_1, α_2, α_3, and α_4 in terms of square roots of expressions involving the roots θ_1, θ_2, and θ_3 of the resolvent cubic. In this case we evidently have

$$(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4) = \theta_1 \qquad (\alpha_1 + \alpha_2) + (\alpha_3 + \alpha_4) = 0$$

which gives

$$\alpha_1 + \alpha_2 = \sqrt{-\theta_1} \qquad \alpha_3 + \alpha_4 = -\sqrt{-\theta_1}.$$

Similarly,

$$\alpha_1 + \alpha_3 = \sqrt{-\theta_2} \qquad \alpha_2 + \alpha_4 = -\sqrt{-\theta_2}$$
$$\alpha_1 + \alpha_4 = \sqrt{-\theta_3} \qquad \alpha_2 + \alpha_3 = -\sqrt{-\theta_3}.$$

An easy computation shows that $\sqrt{-\theta_1}\sqrt{-\theta_2}\sqrt{-\theta_3} = -q$, so that the choice of two of the square roots determines the third. Since $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$, if we add the left-hand equations above we obtain $2\alpha_1$, and similarly we may solve for the other roots of $g(y)$. We find

$$\begin{aligned}
2\alpha_1 &= \sqrt{-\theta_1} + \sqrt{-\theta_2} + \sqrt{-\theta_3} \\
2\alpha_2 &= \sqrt{-\theta_1} - \sqrt{-\theta_2} - \sqrt{-\theta_3} \\
2\alpha_3 &= -\sqrt{-\theta_1} + \sqrt{-\theta_2} - \sqrt{-\theta_3} \\
2\alpha_4 &= -\sqrt{-\theta_1} - \sqrt{-\theta_2} + \sqrt{-\theta_3}
\end{aligned}$$

which reduces the solution of the quartic equation to the solution of the associated resolvent cubic.

EXERCISES

1. Use Cardano's formulas to solve the equation $x^3 + x^2 - 2 = 0$. In particular show that the equation has the real root

$$\frac{1}{3}(\sqrt[3]{26 + 15\sqrt{3}} + \sqrt[3]{26 - 15\sqrt{3}} - 1).$$

Show directly that the roots of this cubic are $1, -1 \pm i$. Explain this by proving that

$$\sqrt[3]{26 + 15\sqrt{3}} = 2 + \sqrt{3} \quad \sqrt[3]{26 - 15\sqrt{3}} = 2 - \sqrt{3}$$

so that

$$\sqrt[3]{26 + 15\sqrt{3}} + \sqrt[3]{26 - 15\sqrt{3}} = 4.$$

2. Let ζ_7 be a primitive 7th root of unity and let $\alpha = \zeta + \zeta^{-1}$.
   (a) Show that ζ_7 is a root of the quadratic $z^2 - \alpha z + 1$ over $\mathbb{Q}(\alpha)$.
   (b) Show using the minimal polynomial for ζ_7 that α is a root of the cubic $x^3 + x^2 - 2x - 1$.
   (c) Use (a) and (b) together with the explicit solution of the cubic in (b) in the text to express ζ_7 in terms of radicals similar to the expressions given earlier for the other roots of unity of small order. (The complicated nature of the expression explains why we did not include ζ_7 earlier in our list of explicit roots of unity.)
3. Let F be a field of characteristic $\neq 2$. State and prove a necessary and sufficient condition on $\alpha, \beta \in F$ so that $F(\sqrt{\alpha}) = F(\sqrt{\beta})$. Use this to determine whether $\mathbb{Q}(\sqrt{1 - \sqrt{2}}) = \mathbb{Q}(i, \sqrt{2})$.
4. Let $K = \mathbb{Q}(\sqrt[n]{a})$, where $a \in \mathbb{Q}$, $a > 0$ and suppose $[K : \mathbb{Q}] = n$ (i.e. $x^n - a$ is irreducible). Let E be any subfield of K and let $[E : \mathbb{Q}] = d$. Prove that $E = \mathbb{Q}(\sqrt[d]{a})$. [Consider $\mathrm{N}_{K/E}(\sqrt[n]{a}) \in E$.]
5. Let K be as in the previous exercise. Prove that if n is odd then K has no nontrivial subfields which are Galois over $\mathbb{Q}$ and if n is even then the only nontrivial subfield of K which is Galois over $\mathbb{Q}$ is $\mathbb{Q}(\sqrt{a})$.

6. Let L be the Galois closure of K in the previous two exercises (i.e. the splitting field of $x^n - a$). Prove that $[L : \mathbb{Q}] = n\varphi(n)$ or $\frac{1}{2}n\varphi(n)$. [Note that $\mathbb{Q}(\zeta_n) \cap K$ is a Galois extension of $\mathbb{Q}$.]

7. (*Kummer generators for cyclic extensions*) Let F be a field of characteristic not dividing n containing the nth roots of unity and let K be a cyclic extension of degree d dividing n. Then $K = F(\sqrt[n]{a})$ for some nonzero $a \in F$. Let σ be a generator for the cyclic group $\mathrm{Gal}(K/F)$.
   (a) Show that $\sigma(\sqrt[n]{a}) = \zeta\sqrt[n]{a}$ for some primitive d^{th} root of unity ζ.
   (b) Suppose $K = F(\sqrt[n]{a}) = F(\sqrt[n]{b})$. Use (a) to show that $\dfrac{\sigma(\sqrt[n]{a})}{\sqrt[n]{a}} = (\dfrac{\sigma(\sqrt[n]{b})}{\sqrt[n]{b}})^i$ for some integer i relatively prime to d. Conclude that σ fixes the element $\dfrac{\sqrt[n]{a}}{(\sqrt[n]{b})^i}$ so this is an element of F.
   (c) Prove that $K = F(\sqrt[n]{a}) = F(\sqrt[n]{b})$ if and only if $a = b^i c^n$ and $b = a^j d^n$ for some $c, d \in F$, i.e. if and only if a and b generate the same subgroup of $F^\times$ modulo n^{th} powers.

8. Let p, q and r be primes in $\mathbb{Z}$ with $q \neq r$. Let $\sqrt[p]{q}$ denote any root of $x^p - q$ and let $\sqrt[p]{r}$ denote any root of $x^p - r$. Prove that $\mathbb{Q}(\sqrt[p]{q}) \neq \mathbb{Q}(\sqrt[p]{r})$.

9. (*Artin–Schreier Extensions*) Let F be a field of characteristic p and let K be a cyclic extension of F of degree p. Prove that $K = F(\alpha)$ where α is a root of the polynomial $x^p - x - a$ for some $a \in F$. [Note that $\mathrm{Tr}_{K/F}(-1) = 0$ since F is of characteristic p so that $-1 = \alpha - \sigma\alpha$ for some $\alpha \in K$ where σ is a generator of $\mathrm{Gal}(K/F)$ by Exercise 26 of Section 2. Show that $a = \alpha^p - \alpha$ is an element of F.] Note that since F contains the pth roots of unity (namely, 1) that this completes the description of all cyclic extensions of prime degree p over fields containing the pth roots of unity in all characteristics.

10. Let $K = \mathbb{Q}(\zeta_p)$ be the cyclotomic field of pth roots of unity for the prime p and let $G = \mathrm{Gal}(K/\mathbb{Q})$. Let ζ denote any p^{th} root of unity. Prove that $\sum_{\sigma \in G} \sigma(\zeta)$ (the trace from K to $\mathbb{Q}$ of ζ) is -1 or $p - 1$ depending on whether ζ is or is not a primitive p^{th} root of unity.

11. (*The Classical Gauss Sum*) Let $K = \mathbb{Q}(\zeta_p)$ be the cyclotomic field of pth roots of unity for the odd prime p, viewed as a subfield of $\mathbb{C}$, and let $G = \mathrm{Gal}(K/\mathbb{Q})$. Let H denote the subgroup of index 2 in the cyclic group G. Define $\eta_0 = \sum_{\tau \in H} \tau(\zeta_p)$, $\eta_1 = \sum_{\tau \in \sigma H} \tau(\zeta_p)$, where σ is a generator of $\mathrm{Gal}(K/\mathbb{Q})$ (the two *periods* of ζ_p with respect to H, i.e. the sum of the conjugates of ζ_p with respect to the two cosets of H in G, cf. Section 5).
   (a) Prove that $\sigma(\eta_0) = \eta_1$, $\sigma(\eta_1) = \eta_0$ and that
   $$\eta_0 = \sum_{a=\text{square}} \zeta_p{}^a \quad , \quad \eta_1 = \sum_{b \neq \text{square}} \zeta_p{}^b \quad ,$$
   where the sums are over the squares and nonsquares (respectively) in $(\mathbb{Z}/p\mathbb{Z})^\times$. [Observe that H is the subgroup of squares in $(\mathbb{Z}/p\mathbb{Z})^\times$.]
   (b) Prove that $\eta_0 + \eta_1 = (\zeta_p, 1) = -1$ and $\eta_0 - \eta_1 = (\zeta_p, -1)$ where $(\zeta_p, 1)$ and $(\zeta_p, -1)$ are two of the Lagrange resolvents of ζ_p.
   (c) Let $g = \sum_{i=0}^{p-1} \zeta_p{}^{i^2}$ (the classical *Gauss sum*). Prove that $g = (\zeta_p, -1) = \sum_{i=0}^{p-2} (-1)^i \sigma^i(\zeta_p)$.
   (d) Prove that $\tau g = g$ if $\tau \in H$ and $\tau g = -g$ if $\tau \notin H$. Conclude in particular that $[\mathbb{Q}(g) : \mathbb{Q}] = 2$. Recall that complex conjugation is the automorphism σ_{-1} on K (cf. Exercise 7 of Section 5). Conclude that $\bar{g} = g$ if -1 is a square mod p (i.e. if

$p \equiv 1 \bmod 4$) and $\bar{g} = -g$ if -1 is not a square mod p (i.e. if $p \equiv 3 \bmod 4$) where $\bar{g}$ denotes the complex conjugate of g.

(e) Prove that $g\bar{g} = p$. [The complex conjugate of a root of unity is its reciprocal. Then $\bar{g} = \sum_{j=0}^{p-2}(-1)^j\,(\sigma^j(\zeta_p))^{-1}$ gives

$$g\bar{g} = \sum_{i,j=0}^{p-2}(-1)^i(-1)^j\frac{\sigma^i(\zeta_p)}{\sigma^j(\zeta_p)} = \sum_{i,j=0}^{p-2}(-1)^{i-j}\sigma^j\left[\frac{\sigma^{i-j}(\zeta_p)}{\zeta_p}\right]$$
$$= \sum_{k=0}^{p-2}(-1)^k\sum_{j=0}^{p-2}\sigma^j\left[\frac{\sigma^k(\zeta_p)}{\zeta_p}\right]$$

where $k = i - j$. If $k = 0$ the element $\dfrac{\sigma^k(\zeta_p)}{\zeta_p}$ is 1 and if $k \neq 0$ then this is a primitive p^{th} root of unity. Use the previous exercise to conclude that the inner sum is $p-1$ when $k = 0$ and is -1 otherwise.]

(f) Conclude that $g^2 = (-1)^{(p-1)/2}p$ and that $\mathbb{Q}\left(\sqrt{(-1)^{(p-1)/2}p}\right)$ is the unique quadratic subfield of $\mathbb{Q}(\zeta_p)$. (cf. also Exercise 33 of Section 6).

12. Let L be the Galois closure of the finite extension $\mathbb{Q}(\alpha)$ of $\mathbb{Q}$. For any prime p dividing the order of $\mathrm{Gal}(L/\mathbb{Q})$ prove there is a subfield F of L with $[L : F] = p$ and $L = F(\alpha)$.

13. Let F be a subfield of the real numbers $\mathbb{R}$. Let a be an element of F and let $K = F(\sqrt[n]{a})$ where $\sqrt[n]{a}$ denotes a real n^{th} root of a. Prove that if K is Galois over F then $[K : F] \leq 2$.

14. This exercise shows that in general it is necessary to use complex numbers when expressing real roots in terms of radicals and generalizes the *Casus irreducibilis* of cubic equations. Let $f(x) \in \mathbb{Q}[x]$ be an irreducible polynomial all of whose roots are real. Suppose further that one of the roots, α, of $f(x)$ can be expressed in terms of *real* radicals (i.e. there is a root extension of real fields: $\mathbb{Q} = K_0 \subset K_1 \subset \ldots \subset K_m \subset \mathbb{R}$ with $K_{i+1} = K_i(\sqrt[n]{a_i})$, $i = 1, 2, \ldots, m-1$, for some integers n_i and some $a_i \in K_i$ and $\alpha \in K_m$). Prove that the Galois group of $f(x)$ is a 2–group. Conclude in particular that the degree of $f(x)$ is a power of 2 and that the real roots of such a polynomial can be expressed entirely in terms of real radicals if and only if these roots can be constructed by straightedge and compass. [The argument is similar to the case of cubics. Let $L \in \mathbb{R}$ be the Galois closure of $\mathbb{Q}(\alpha)$ and suppose the order of $\mathrm{Gal}(L/\mathbb{Q})$ is divisible by some odd prime p. Let F be a subfield of L with $[L : F] = p$ and $L = F(\alpha)$ (by Exercise 12) and consider the composite fields $K_i' = FK_i$, $i = 0, 1, \ldots, m$. These are again real radical extensions and by the argument in the text for the *Casus irreducibilis*, we may assume each $[K_{i+1}' : K_i']$ is a prime. Since $\alpha \notin F = FK_0$, there is an integer s with $\alpha \notin K_{s-1}'$, $\alpha \in K_s'$. Since the extensions are of prime degree, we have $K_s' = K_{s-1}'(\alpha)$. Since $L = F(\alpha)$ is Galois of degree p, K_s' is a Galois extension of K_{s-1}' of degree p, contradicting the previous exercise.]

15. Let F be a subfield of the real numbers $\mathbb{R}$. Let a be an element of F and let $K = F(\sqrt[n]{a})$ where $\sqrt[n]{a}$ denotes a real n^{th} root of a. Prove that if L is any Galois extension of F contained in K then $[L : F] \leq 2$.

16. Let a be a nonzero rational number.

(a) Determine when the extension $\mathbb{Q}(\sqrt{ai})$ ($i^2 = -1$) is of degree 4 over $\mathbb{Q}$.

(b) When $K = \mathbb{Q}(\sqrt{ai})$ is of degree 4 over $\mathbb{Q}$ show that K is Galois over $\mathbb{Q}$ with the Klein 4–group as Galois group. In this case determine the quadratic extensions of $\mathbb{Q}$ contained in K.

17. Let $D \in \mathbb{Z}$ be a squarefree integer and let $a \in \mathbb{Q}$ be a nonzero rational number. Show that $\mathbb{Q}(\sqrt{a\sqrt{D}})$ cannot be a cyclic extension of degree 4 over $\mathbb{Q}$.

18. Let $D \in \mathbb{Z}$ be a squarefree integer and let $a \in \mathbb{Q}$ be a nonzero rational number. Prove that if $\mathbb{Q}(\sqrt{a\sqrt{D}})$ is Galois over $\mathbb{Q}$ then $D = -1$.

19. Let $D \in \mathbb{Z}$ be a squarefree integer and let $K = \mathbb{Q}(\sqrt{D})$.
    (a) Prove that if $D = s^2 + t^2$ is the sum of two rational squares then there exists an extension $L/\mathbb{Q}$ containing K which is Galois over $\mathbb{Q}$ with a cyclic Galois group of order 4. [Consider the extension $\mathbb{Q}(\sqrt{D + s\sqrt{D}})$.] (Note also that D is the sum of two rational squares if and only if D is also the sum of two integer squares, so one may assume s and t are integral without loss.)
    (b) Prove conversely that if K can be embedded in a cyclic extension L of degree 4 as in (a) then D is the sum of two squares. [One approach: (i) observe first that L is quadratic over K, so $L = K(\sqrt{a + b\sqrt{D}})$ for some $a, b \in \mathbb{Q}$, (ii) show that L contains the quadratic subfield $\mathbb{Q}(\sqrt{a^2 - b^2 D})$, which must be $\mathbb{Q}(\sqrt{D})$ if $L/\mathbb{Q}$ is cyclic, and use Exercise 7.]
    (c) Conclude in particular that $\mathbb{Q}(\sqrt{3})$ is not a subfield of any cyclic extension of degree 4 over $\mathbb{Q}$. Similarly conclude that the fields $\mathbb{Q}(\sqrt{D})$ for squarefree integers $D < 0$ are never contained in cyclic extensions of degree 4 over $\mathbb{Q}$ (this gives an alternate proof for Exercise 19, Section 6).

20. Let p be a prime. Show that any solvable subgroup of S_p of order divisible by p is contained in the normalizer of a Sylow p-subgroup of S_p (a Frobenius group of order $p(p-1)$). Conclude that an irreducible polynomial $f(x) \in \mathbb{Q}[x]$ of degree p is solvable by radicals if and only if its Galois group is contained in the Frobenius group of order $p(p-1)$. [Let $G \leq S_p$ be a solvable subgroup of order divisible by p. Then G contains a p-cycle, hence is transitive on $\{1, 2, \ldots, p\}$. Let $H < G$ be the stabilizer in G of the element 1, so H has index p in G. Show that H contains no nontrivial normal subgroups of G (note that the conjugates of H are the stabilizers of the other points). Let $G^{(n-1)}$ be the last nontrivial subgroup in the derived series for G. Show that $H \cap G^{(n-1)} = 1$ and conclude that $|G^{(n-1)}| = p$, so that the Sylow p-subgoup of G (which is also a Sylow p-subgroup in S_p) is normal in G.]

21. (*Criterion for the Solvability of a Quintic*) By the previous exercise, an irreducible polynomial $f(x)$ in $\mathbb{Q}[x]$ of degree 5 can be solved by radicals if and only if its Galois group (considered as a subgroup of S_5) is contained in the Frobenius group of order 20. It is known that this is the case if and only if an associated polynomial $g(x)$ of degree 6 has a rational root. If the quintic is in the general form (where a translation is performed so that the coefficient of x^4 is zero):

$$f(x) = x^5 + px^3 + qx^2 + rx + s \qquad p, q, r, s \in \mathbb{Q}$$

then the associated polynomial of degree 6 is:

$$\begin{aligned} g(x) = x^6 &+ 8rx^5 + (2pq^2 - 6p^2r + 40r^2 - 50qs)\,x^4 \\ &+ (-2q^4 + 21pq^2r - 40p^2r^2 + 160r^3 - 15p^2qs - 400qrs + 125ps^2)\,x^3 \\ &+ (p^2q^4 - 8q^4r + 9p^4r^2 - 136p^2r^3 + 625q^2s^2 + 400r^4 - 6p^3q^2r \\ &\quad + 76pq^2r^2 - 50pq^3s - 1400qr^2s + 500prs^2 + 90p^2qrs)\,x^2 \\ &+ (-108p^5s^2 + 32p^4r^3 - 256p^2r^4 - 3125s^4 + 512r^5 - 2pq^6 + 3q^4r^2 \\ &\quad - 58q^5s + 2750q^2rs^2 - 31p^3q^3s - 500pr^2s^2 + 19p^2q^4r \\ &\quad - 51p^3q^2r^2 + 76pq^2r^3 - 2400qr^3s - 325p^2q^2s^2 + 525p^3rs^2 \end{aligned}$$

$$
\begin{aligned}
&+ 625pqs^3 + 117p^4qrs + 105pq^3rs + 260p^2qr^2s)\,x \\
&+ (q^8 + 256r^6 + 17q^4r^3 - 27p^7s^2 - 4p^6r^3 + 48p^4r^4 - 192p^2r^5 \\
&+ 3125p^2s^4 - 9375rs^4 - 1600qr^4s - 99p^5rs^2 - 125pq^4s^2 \\
&- 124q^5rs + 3250q^2r^2s^2 - 2000pr^3s^2 - 13pq^6r + p^5q^2r^2 \\
&+ 65p^2q^4r^2 - 128p^3q^2r^3 - 16pq^2r^4 - 4p^5q^3s - 12p^2q^5s \\
&- 150p^4q^2s^2 + 1200p^3r^2s^2 + 18p^6qrs + 12p^3q^3rs + 196p^4qr^2s \\
&+ 590pq^3r^2s - 160p^2qr^3s - 725p^2q^2rs^2 - 1250pqrs^3).
\end{aligned}
$$

In the particular case where $f(x) = x^5 + Ax + B$ this polynomial is simply

$$g(x) = x^6+8Ax^5+40A^2x^4+160A^3x^3+400A^4x^2+(512A^5-3125B^4)x-9375AB^4+256A^6$$

(a) Use this criterion to prove that the Galois group over $\mathbb{Q}$ of the polynomial $x^5 - 5x + 12$ is the dihedral group of order 10. [Show the associated sixth degree polynomial is

$$x^6 - 40x^5 + 1000x^4 - 20000x^3 + 250000x^2 - 66400000x + 976000000$$

and has $x = 40$ as a rational root.]

(b) Use this criterion to prove that $x^5 - x - 1$ is not solvable by radicals.

14.8 COMPUTATION OF GALOIS GROUPS OVER $\mathbb{Q}$

In the determination of the Galois groups of polynomials of degrees ≤ 4 in Section 6 and in the determination of the Galois group of the polynomial $x^5 - 6x + 3$ in the previous section we observed that it was possible to obtain useful information regarding the Galois group from the *cycle types* of the automorphisms as elements in S_n. This is very useful in computing Galois groups of polynomials over $\mathbb{Q}$ and we now briefly describe the theoretical justification.

Let $f(x)$ be a polynomial with rational coefficients. In determining the Galois group of $f(x)$ we may assume that $f(x)$ is separable and has integer coefficients. Then the discriminant D of $f(x)$ is an integer and is nonzero.

For any prime p, consider the reduction $\bar{f}(x) \in \mathbb{F}_p[x]$ of $f(x)$ modulo p. If p divides D then the reduced polynomial $\bar{f}(x)$ has discriminant $\bar{D} = 0$ in $\mathbb{F}_p$, so is not separable.

If p does not divide D, then $\bar{f}(x)$ is a separable polynomial over $\mathbb{F}_p$ and we can factor $\bar{f}(x)$ into distinct irreducibles

$$\bar{f}(x) = \bar{f}_1(x)\bar{f}_2(x)...\bar{f}_k(x) \qquad \text{in } \mathbb{F}_p[x].$$

Let n_i be the degree of $\bar{f}_i(x)$, $i = 1, 2, ..., k$.

The importance of this reduction is provided by the following theorem from algebraic number theory which is an elementary consequence of the study of the arithmetic in finite extensions of $\mathbb{Q}$ (and which we take for granted):

Theorem. For any prime p not dividing the discriminant D of $f(x) \in \mathbb{Z}[x]$, the Galois group over $\mathbb{F}_p$ of the reduction $\bar{f}(x) = f(x) \pmod p$ is permutation group isomorphic to a subgroup of the Galois group over $\mathbb{Q}$ of $f(x)$.

The meaning of the statement "permutation group isomorphic" in the theorem is that not only is the Galois group of the reduction $\bar{f}(x)$ mod p of $f(x)$ isomorphic to a subgroup of the Galois group of $f(x)$ but that there is an ordering of the roots of $\bar{f}(x)$ and of $f(x)$ (depending on p) so that under this isomorphism the action of the corresponding automorphisms as permutations of these roots is the same. In particular there are automorphisms in the Galois group of $f(x)$ with the same cycle types as the automorphisms of $\bar{f}(x)$.

The Galois group of $\bar{f}(x)$ is a *cyclic* group since every finite extension of $\mathbb{F}_p$ is a cyclic extension. Let σ be a generator for this Galois group over $\mathbb{F}_p$ (for example, the Frobenius automorphism). The roots of $\bar{f}_1(x)$ are permuted amongst themselves by the Galois group and given any two of these roots there is a Galois automorphism taking the first root to the second (recall that the group is said to be *transitive* on the roots when this is the case). Similarly, the Galois group permutes the roots of each of the factors $\bar{f}_i(x)$, $i = 1, 2, ..., k$ transitively. Since these factors are relatively prime we also see that no root of one factor is mapped to a root of any other factor by any element of the Galois group.

View σ as an element in S_n by labelling the n roots of $\bar{f}(x)$ and consider the cycle decomposition of σ, which is a product of k distinct permutations since σ permutes the roots of each of the factors $\bar{f}_i(x)$ amongst themselves. By the observations we just made, the action of σ on the roots of $\bar{f}_1(x)$ must be a cycle of length n_i since otherwise the powers of σ could not be transitive on the roots of $\bar{f}_1(x)$. Similarly the action of σ on the roots of $\bar{f}_i(x)$ gives a cycle of length n_i, $i = 1, 2, ..., k$.

We see that the automorphism σ generating the Galois group of $\bar{f}(x)$ has cycle decomposition $(n_1, n_2, ..., n_k)$ where $n_1, n_2, ..., n_k$ are the degrees of the irreducible factors of $f(x)$ reduced modulo p, which gives us

Corollary 41. For any prime p not dividing the discriminant of $f(x) \in \mathbb{Z}[x]$, the Galois group of $f(x)$ over $\mathbb{Q}$ contains an element with cycle decomposition $(n_1, n_2, ..., n_k)$ where $n_1, n_2, ..., n_k$ are the degrees of the irreducible factors of $f(x)$ reduced modulo p.

Example

Consider the polynomial $x^5 - x - 1$. The discriminant of this polynomial is $2869 = 19 \cdot 151$ so we reduce at primes $\neq 19, 151$. Reducing mod 2 we have $x^5 - x - 1 \equiv (x^2 + x + 1)(x^3 + x^2 + 1) \pmod 2$ so the Galois group has a (2,3)-cycle. Cubing this element we see the Galois group contains a transposition.

Reducing mod 3 the polynomial is irreducible, as follows: $x^5 - x - 1$ has no roots mod 3 so if it were reducible mod 3 then it would have an irreducible quadratic factor, hence would have a factor in common with $x^9 - x$ (which is the product of all irreducible polynomials of degrees 1 and 2 over $\mathbb{F}_3$), hence a factor in common with either $x^4 - 1$ or $x^4 + 1$, hence a factor in common with either $x^5 - x$ or $x^5 + x$, hence a factor in common with either -1 or $2x + 1$ which it obviously does not. This shows both that $x^5 - x - 1$ is irreducible in $\mathbb{Z}[x]$ and that there is a 5-cycle in its Galois group.

Since S_5 is generated by any 5-cycle and any transposition, it follows that the Galois group of $x^5 - x - 1$ is S_5 (so in particular this polynomial cannot be solved by radicals, (cf. Exercise 21 of Section 7).

The arguments in the example above indicate how to construct polynomials with S_n as Galois group. We use the fact that a transitive subgroup of S_n containing a transposition and an $n-1$-cycle is S_n. Let f_1 be an irreducible polynomial of degree n over $\mathbb{F}_2$. Let $f_2 \in \mathbb{F}_3[x]$ be the product of an irreducible polynomial of degree 2 with irreducible polynomials of odd degree (for example, an irreducible polynomial of degree $n-3$ and x if n is even and an irreducible polynomial of degree $n-2$ if n is odd). Let $f_3 \in \mathbb{F}_5[x]$ be the product of x with an irreducible polynomial of degree $n-1$. Finally, let $f(x) \in \mathbb{Z}[x]$ be any polynomial with

$$\begin{aligned} f(x) &\equiv f_1(x) \pmod 2 \\ &\equiv f_2(x) \pmod 3 \\ &\equiv f_3(x) \pmod 5 \ . \end{aligned}$$

The reduction of $f(x)$ mod 2 shows that $f(x)$ is irreducible in $\mathbb{Z}[x]$, hence the Galois group is transitive on the n roots of $f(x)$. Raising the element given by the factorization of $f(x)$ mod 3 to a suitable odd power shows the Galois group contains a transposition. The factorization mod 5 shows the Galois group contains an $n-1$-cycle, hence the Galois group is S_n.

Proposition 42. There exist infinitely many polynomials $f(x) \in \mathbb{Z}[x]$ with S_n as Galois group.

There are extremely efficient algorithms for factoring polynomials $f(x) \in \mathbb{Z}[x]$ modulo p (cf. Exercises 12 to 17 of Section 3), so the corollary above is an effective procedure for determining some of the cycle types of the elements of the Galois group. In using this corollary, some care should be taken not to assume that a *particular* cycle is an element of the Galois group. For example, one factorization might imply the existence of a (2,2) cycle, say (12)(34) and another factorization imply the existence of a transposition. One cannot conclude that the transposition is necessarily (12), however (nor (34), nor (13), etc.). The choice of (12)(34) to represent the first cycle fixes a particular ordering on the roots and this may not be the ordering with respect to which the transposition appears as (12).

This corollary is particularly efficient in determining when the Galois group is large (e.g. S_n), since a transitive group containing sufficiently many cycle types must be S_n (for example, a transitive subgroup of S_n containing a transposition and an $n-1$-cycle is S_n, as used above). The most difficult Galois groups to determine in this way are the *small* Galois groups (e.g. a cyclic group of order n), since factorization after factorization will produce only elements of orders dividing n and one is not sure whether there will be some p yet to come producing a cycle type inconsistent with the assumption of a cyclic Galois group. If one could "compute forever" one could at least be sure of the precise distribution of cycle types among the elements of the Galois group in the following sense: suppose the Galois group $G \subseteq S_n$ has order N and that there are n_T elements of G with cycle type T (e.g. (2,2)-cycles, transpositions,...) so that the "density" of cycle type T in G is $d_T = n_T/N$. Then it is possible to define a density on the set of prime numbers (so that it makes sense to speak of "1/2"

the primes, etc.) and we have the following result (which relies on the Tchebotarov Density Theorem in algebraic number theory):

Theorem. The density of primes p for which $f(x)$ splits into type T modulo p is precisely d_T.

This says that if we knew the factorization of $f(x)$ modulo every prime we could at least determine the number of elements of G with a given cycle type. Unfortunately, even this would not be sufficient to determine G (up to isomorphism): it is known that there are nonisomorphic groups containing the same number of elements of all cycle types (there are two nonisomorphic groups of order 96 in S_8 both having cycle type distributions: 1 1-cycle, 6 (2,2)-cycles, 13 (2,2,2,2)-cycles, 32 (3,3)-cycles, 12 (4,4)-cycles, 32 (2,6)-cycles). There are infinitely many such examples (the regular representation of the elementary abelian group of order p^3 and for the nonabelian group of order p^3 of exponent p give two nonisomorphic groups in S_{p^3} whose nonidentity elements are all the product of p^2 p-cycles for any prime p).

In practice one uses the factorizations of $f(x)$ modulo small primes to get an idea of the probable Galois group (based on the previous result). One then tries to prove this is indeed the Galois group – often a difficult problem. For polynomials of small degree, definitive algorithms exist, based in part on the computation of *resolvent* polynomials. These are analogues of the cubic resolvent used in the previous sections to determine the Galois group of quartic polynomials. These resolvent polynomials have rational coefficients and have as roots certain combinations of the roots of $f(x)$ (similar to the combinations $(\alpha_1 + \alpha_2)(\alpha_3 + \alpha_4)$ for the cubic resolvent). One then determines the factorization of these resolvent polynomials to obtain information on the Galois group of $f(x)$ – for example the existence of a linear factor implies the Galois group lies in the stabilizer in S_n of the combination of the roots of $f(x)$ chosen (for example, the dihedral group of order 8 for our resolvent cubic). It should be observed, however, that the degree of the resolvent polynomials constructed, unlike the situation of the resolvent cubic for quartic polynomials, are in general much larger than the degree of $f(x)$. The effectiveness of this computational technique also depends heavily on the explicit knowledge of the possible transitive subgroups of S_n. For $n = 2, 3, \ldots, 8$ the number of isomorphism classes of transitive subgroups of S_n is $1, 2, 5, 5, 16, 7, 50$, respectively. There is a good deal of interest in the computation of Galois groups, motivated in part by the problem of determining which groups occur as Galois groups over $\mathbb{Q}$.

We illustrate these techniques with some easier examples.[6]

Examples

(1) There are 5 isomorphism classes of transitive subgroups of S_5 given by the groups Z_5, D_{10}, F_{20}, the so-called Frobenius group of order 20 (the Galois group of $x^5 - 2$ with

[6] These examples are taken from *The Computation of Galois Groups*, L. Soicher, Master's Thesis, Concordia University, Montreal, 1981. See also Computing Galois Groups over the Rationals, L. Soicher and J. McKay, *Jour. of Number Theory*, 20, pp. 273–281, (1985).

generators $(1\,2\,3\,4\,5)$ and $(2\,3\,5\,4)$ in S_5), A_5 and S_5. The cycle type distributions for these groups are as follows:

cycle type :	1	2	(2,2)	3	(2,3)	4	5
Z_5	1						4
D_{10}	1		5				4
F_{20}	1		5			10	4
A_5	1		15	20			24
S_5	1	10	15	20	20	30	24.

Given this information, the irreducibility of $x^5 - x - 1$ (giving the transitivity on the 5 roots) and the cycle type (2,3) immediately shows that the Galois group of $x^5 - x - 1$ is S_5.

Consider now the polynomial $x^5 + 15x + 12$. The discriminant is $2^{10}3^4 5^5$ so the Galois group is not contained in A_5. There are two possibilities: S_5 or F_{20}. One can easily determine which is more likely by factoring the polynomial modulo a number of small primes and comparing the distribution of cycle types with those in the table above. This does not *prove* the probable Galois group is actually correct. To decide which of S_5 and F_{20} is correct one can compute the resolvent polynomial $R(x)$ of degree 15 whose roots are the distinct permutations under S_5 of $(\alpha_1 + \alpha_2 - \alpha_3 - \alpha_4)^2$ for 4 of the roots $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ of $f(x)$. By definition, S_5 is transitive on the roots of $R(x)$ and it is not difficult to check using the explicit generators for F_{20} given above that F_{20} is not transitive on these 15 values. It follows that $R(x)$ will be a reducible polynomial over $\mathbb{Q}$ if and only if the Galois group of the quintic is F_{20}. One finds that for $x^5 + 15x + 12$ the resolvent polynomial $R(x)$ factors into a polynomial of degree 5 and a polynomial of degree 10, hence the Galois group for this quintic is F_{20}. One can also use Exercise 21 of the previous section (cf. Exercise 6), which is also based on the computation of a related resolvent polynomial.

(2) Consider the polynomial $x^7 - 14x^5 + 56x^3 - 56x + 22$. The discriminant is computed to be $2^6 7^{10}$ so the Galois group is contained in A_7.

Factoring the polynomial for the 42 primes not equal to 7 between 3 and 193 gives a cycle type distribution of 1 1-cycle (2.38 %), 30 (3,3)-cycles (71.43 %), 11 7-cycles (26.19 %). There are 7 isomorphism classes of transitive subgroups of S_7, 4 of them contained in A_7. Of these, one contains no (3,3)-cycles, which leaves the three possibilities A_7, $GL_3(\mathbb{F}_2)$, and F_{21}, the Frobenius group of order 21 (which has generators $(1\,2\,3\,4\,5\,6\,7)$ and $(2\,3\,5)(4\,7\,6)$ in S_7). The cycle type distributions for these three are as follows:

cycle type :	1	2	(2,2)	3	(2,2,3)	(3,3)	(2,4)	5	7
F_{21}	1					14			6
$GL_3(\mathbb{F}_2)$	1		21			56	42		48
A_7	1	21	105	70	210	280	630	504	720

It follows that there is a strong probability that the Galois group of this polynomial is the Frobenius group of order 21. This is actually the case (the verification requires computation of a resolvent of degree 35 and factoring it over $\mathbb{Z}$ – there are three factors, of degrees 7,7, and 21).

EXERCISES

1. Let p be a prime. Prove that the polynomial $x^4 + 1$ splits mod p either into two irreducible quadratics or into 4 linear factors using Corollary 41 together with the knowledge that the Galois group of this polynomial is the Klein 4-group.
2. (cf. Exercise 48 of Section 6).
   (a) Let K be the splitting field of $x^6 - 2x^3 - 2$. Prove that if $[K : \mathbb{Q}] = 12$ then $K = \mathbb{Q}(\sqrt[3]{2}, i, \sqrt{3})$ and K is generated over the biquadratic field $F = \mathbb{Q}(i, \sqrt{3})$ by $\alpha = \sqrt[3]{1+\sqrt{3}}$ and by $\beta = \sqrt[3]{1-\sqrt{3}}$. Show that if this is the case then the elements of order 3 in $\mathrm{Gal}(K/\mathbb{Q})$ lie in $\mathrm{Gal}(K/F)$. Conclude that any element of $\mathrm{Gal}(K/\mathbb{Q})$ of order 3 maps α to another cube root of $1+\sqrt{3}$ and maps β to another cube root of $1-\sqrt{3}$ and if it is the identity on α or β then it is the identity on all of K.
   (b) Show that the factorization of $f(x)$ into irreducibles over $\mathbb{F}_{13}$ is $(x-7)(x-8)(x-11)(x^3+3)$ and use Corollary 41 to show that $[K : \mathbb{Q}] = 36$.
   (c) Knowing that $G = \mathrm{Gal}(K/\mathbb{Q})$ is of order 36 determine all the elements of G explicitly and in particular show that G is isomorphic to $S_3 \times S_3$.
3. Prove that the Galois group of $x^5 + 20x + 16$ is A_5.
4. Prove that the Galois group of $x^5 + x^4 - 4x^3 - 3x^2 + 3x + 1$ is cyclic of order 5. [Show this is the minimal polynomial of $\zeta_{11} + \zeta_{11}^{-1}$.]
5. Prove that the Galois group of $x^5 + 11x + 44$ is the dihedral group D_{10} (cf. Exercise 21 of Section 7).
6. Prove that the Galois group of $x^5 + 15x + 12$ is F_{20}, the Frobenius group of order 20 (cf. Exercise 21 of Section 7).
7. Prove that the Galois group of $x^6 + 24x - 20$ is A_6.
8. Prove that the Galois group of $x^7 + 7x^4 + 14x + 3$ is A_7.
9. Determine a polynomial of degree 7 whose Galois group is cyclic of order 7.
10. Determine the probable Galois group of $x^7 - 7x + 3$.

14.9 TRANSCENDENTAL EXTENSIONS, INSEPARABLE EXTENSIONS, INFINITE GALOIS GROUPS

This section collects some results on arbitrary extensions E/F. These results supplement those of the preceding sections and complete the basic picture of how an arbitrary (possibly infinite) extension decomposes. Since this section is primarily intended as a survey, none of the proofs are included; whenever these proofs can be easily supplied by the reader we indicate this either in the text or (with hints) in the exercises.

Throughout this section E/F is an extension of fields. Recall that an element of E which is not algebraic over F is called transcendental over F. Keep in mind that extensions involving transcendentals are always of infinite degree. We generally reserve the expression: t is an "indeterminate" over F, when we are thinking of evaluating t. Field theoretically, however, the terms transcendental and indeterminate are synonymous (so that the subfield $\mathbb{Q}(\pi)$ of $\mathbb{R}$ and the field $\mathbb{Q}(t)$ are isomorphic).

Definition

(1) A subset $\{a_1, a_2, \ldots, a_n\}$ of E is called *algebraically independent* over F if there is no nonzero polynomial $f(x_1, x_2, \ldots, x_n) \in F[x_1, x_2, \ldots, x_n]$ such that $f(a_1, a_2, \ldots, a_n) = 0$. An arbitrary subset S of E is called *algebraically independent* over F if every finite subset of S is algebraically independent. The elements of S are called *independent transcendentals* over F.

(2) A *transcendence base* for E/F is a maximal subset (with respect to inclusion) of E which is algebraically independent over F.

Note that if E/F is algebraic, the empty set is the only algebraically independent subset of E. In particular, elements of an algebraically independent set are necessarily transcendental. Moreover, one easily checks that $S \subseteq E$ is an algebraically independent set over F if and only if each $s \in S$ is transcendental over $F(S - \{s\})$. It is also an easy exercise to see that S is a transcendence base for E/F if and only if S is a set of algebraically independent transcendentals over F and E is algebraic over $F(S)$.

Theorem. The extension E/F has a transcendence base and any two transcendence bases of E/F have the same cardinality.

Proof: The first statement is a standard Zorn's Lemma argument. The proof of the second uses the same "Replacement Lemma" idea as was used to prove that any two bases of a vector space have the same cardinality.

Definition. The cardinality of a transcendence base for E/F is called the *transcendence degree* of E/F.

Algebraic extensions are precisely the extensions of transcendence degree 0.

One special case of this theorem is when E is *finitely generated* over F, that is, $E = F(\alpha_1, \alpha_2, \ldots, \alpha_n)$, for some (not necessarily algebraically independent) elements $\alpha_1, \ldots, \alpha_n$ of E. It is clear that we may renumber $\alpha_1, \ldots, \alpha_n$ so that $\alpha_1, \ldots, \alpha_m$ are independent transcendentals and $\alpha_{m+1}, \ldots, \alpha_n$ are algebraic over $F(\alpha_1, \ldots, \alpha_m)$ (so E is a finite extension of the latter field). In this case E is called a *function field in m variables* over F. Such fields play a fundamental role in Algebraic Geometry as fields of functions on m-dimensional surfaces. For instance, when $F = \mathbb{C}$ and $m = 1$, these fields arise in analysis as fields of meromorphic functions on compact Riemann surfaces.

Note that if S_1 and S_2 are transcendence bases for E/F it is not necessarily the case that $F(S_1) = F(S_2)$. For example, if t is transcendental over $\mathbb{Q}$, $\{t\}$ and $\{t^2\}$ are both transcendence bases for $\mathbb{Q}(t)/\mathbb{Q}$ but (as we shall see shortly) $\mathbb{Q}(t^2)$ is a proper subfield of $\mathbb{Q}(t)$.

We now see that if $x_1, x_2, \ldots, x_n$ are indeterminates over F and

$$f(x) = (x - x_1)(x - x_2) \cdots (x - x_n) \tag{14.28}$$

is the general polynomial of degree n, then the elementary symmetric functions $s_1, s_2, \ldots, s_n$ in the x_i's are also independent transcendentals over F. This is because

$x_1, \ldots, x_n$ is a transcendence base for $E = F(x_1, \ldots, x_n)$ over F (so the transcendence degree is n) and E is algebraic over $F(s_1, \ldots, s_n)$ (of degree $n!$). The theorem forces $s_1, \ldots, s_n$ to be a transcendence base for this extension as well (in particular, they are independent transcendentals). The general polynomial of degree n over F may therefore equivalently be defined by taking $a_1, \ldots, a_n$ to be any independent transcendentals (or indeterminates) and letting

$$f(x) = x^n + a_1 x^{n-1} + \cdots + a_n \tag{14.29}$$

where the roots of f are denoted by $x_1, \ldots, x_n$ (and $s_i = (-1)^i a_i$).

Definition. An extension E/F is called *purely transcendental* if it has a transcendence base S such that $E = F(S)$.

In the preceding discussion, both $F(x_1, \ldots, x_n)$ and $F(s_1, \ldots, s_n)$ are purely transcendental over F. As an exercise (following) one can show that $\mathbb{Q}(t, \sqrt{t^3 - t})$ is not a purely transcendental extension of $\mathbb{Q}$ even though it contains no elements which are algebraic over $\mathbb{Q}$ other than those in $\mathbb{Q}$ itself (i.e. the process of decomposing a general extension into a purely transcendental extension followed by an algebraic extension cannot generally be reversed so that the algebraic piece occurs first).

If E is a purely transcendental extension of F of transcendence degree $n = 1$ or 2 and L is an intermediate field, $F \subseteq L \subseteq E$ with the same transcendence degree, then L is again a purely transcendental extension of F (Lüroth ($n = 1$), Castelnuovo ($n = 2$)). This result is not true if the transcedence degree is ≥ 3, however, although examples where L fails to be purely transcendental are difficult to construct. For extensions of transcendence degree 1 the intermediate fields are described by the following:

Theorem. Let t be transcendental over F.

(1) (Lüroth) If $F \subseteq K \subseteq F(t)$, then $K = F(r)$, for some $r \in F(t)$. In particular, every nontrivial extension of F contained in $F(t)$ is purely transcendental over F.

(2) If $P = P(t), Q = Q(t)$ are nonzero relatively prime polynomials in $F[t]$ which are not both constant,

$$[F(t) : F(P/Q)] = \max(\deg P, \deg Q).$$

Proof: The proof of (2) is outlined in Exercise 18 of Section 13.2.

By part (2) of this theorem we see that $F(P/Q) = F(t)$ if and only if P, Q are nonzero relatively prime polynomials of degree ≤ 1 (not both constant). Thus $F(r) = F(t)$ if and only if $r = \dfrac{at + b}{ct + d}$, where $a, b, c, d \in F$ and $ad - bc \neq 0$ (called a *fractional linear transformation of* t). For *any* $r \in F(t) - F$ the map $t \mapsto r$ extends to an embedding of $F(t)$ into itself which is the identity on F. This embedding is surjective

(i.e. is an automorphism of $F(t)$) precisely for the fractional linear transformations. Furthermore, the map

$$GL_2(F) \to \mathrm{Aut}(F(t)/F) \quad \text{defined by} \quad \begin{pmatrix} a & c \\ b & d \end{pmatrix} \mapsto (t \mapsto \frac{at+b}{ct+d})$$

is a surjective homomorphism with kernel consisting of the scalar matrices (where $(t \mapsto \frac{at+b}{ct+d})$ denotes the automorphism of $F(t)$ defined by mapping t to $\frac{at+b}{ct+d}$). Thus

$$\mathrm{Aut}(F(t)/F) \cong PGL_2(F)$$

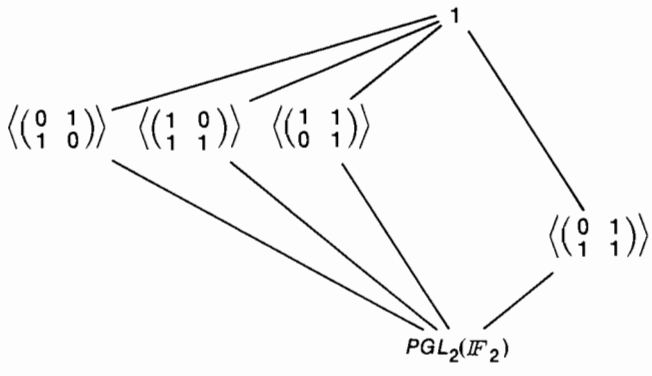
where $PGL_2(F) = GL_2(F)/\{\lambda I \mid \lambda \in F^\times\}$ gives the group of automorphisms of this transcendental extension (cf. Exercise 8 of Section 1).

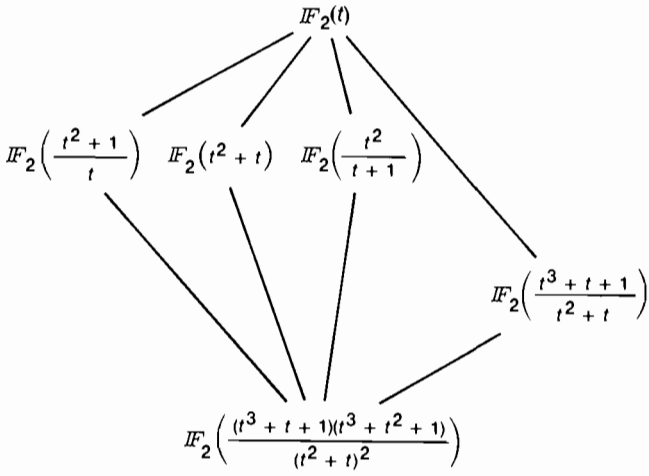
When $\mathbb{F}$ is a finite field of order q, $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F}) \cong PGL_2(\mathbb{F})$ is a finite group of order $q(q-1)(q+1)$. By Corollary 11 if K is the fixed field of $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$, then $\mathbb{F}(t)$ is Galois over K with Galois group equal to $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$. In particular, the fixed field of $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$ is not $\mathbb{F}$ in this case.

This also provides further examples of the Galois correspondence which can be written out completely for small values of q. For instance, if $q = |\mathbb{F}| = 2$, $PGL_2(\mathbb{F})$ is nonabelian of order 6, hence is isomorphic to S_3. The field $\mathbb{F}(t)$ is of degree 6 over the fixed field K of $\mathrm{Aut}(\mathbb{F}(t)/\mathbb{F})$ and the lattice of subfields $K \subseteq L \subseteq \mathbb{F}(t)$ is dual to the lattice of subgroups of S_3. The fixed field of a cyclic subgroup $\langle\sigma\rangle$ is easily found (via the preceding theorem) by finding a rational function r in t which is fixed by σ such that $[\mathbb{F}(t) : \mathbb{F}(r)] = |\sigma|$. For example, if $\sigma : t \mapsto \frac{1}{t+1}$, then σ has order 3. The rational function

$$r = t + \sigma(t) + \sigma^2(t) = \frac{t^3+t+1}{t(t+1)}$$

is fixed by σ and $[\mathbb{F}(t) : \mathbb{F}(r)] = 3$ (by part (2) of the theorem). Since $\mathbb{F}(r)$ is contained in the fixed field of $\langle\sigma\rangle$ and the degree of $\mathbb{F}(t)$ over the fixed field is 3, $\mathbb{F}(r)$ is the fixed field of $\langle\sigma\rangle$. In this way one can explicitly describe the lattice of all subfields of $\mathbb{F}(t)$ containing K:





Purely transcendental extensions of $\mathbb{Q}$ play an important role in the problem of realizing finite groups as Galois groups over $\mathbb{Q}$. We describe a deep result of Hilbert which is fundamental to this area of research. If $a_1, a_2, \ldots, a_n$ are independent indeterminates over a field F, we may evaluate (or *specialize*) $a_1, \ldots, a_n$ at any elements of F, i.e. substitute values in F for the "variables" $a_1, a_2, \ldots, a_n$. If E is a Galois extension of $F(a_1, \ldots, a_n)$, then E is obtained as a splitting field of a polynomial whose coefficients lie in $F[a_1, \ldots, a_n]$. Any specialization of $a_1, \ldots, a_n$ into F maps this polynomial into one whose coefficients lie in F. The specialization of E is the splitting field of the resulting specialized polynomial.

Theorem. Let $x_1, x_2, \ldots, x_n$ be independent transcendentals over $\mathbb{Q}$, let $E = \mathbb{Q}(x_1, \ldots, x_n)$ and let G be a finite group of automorphisms of E with fixed field K. If K is a purely transcendental extension of $\mathbb{Q}$ with transcendence basis $a_1, a_2, \ldots, a_n$, then there are infinitely many specializations of $a_1, \ldots, a_n$ in $\mathbb{Q}$ such that E specializes to a Galois extension of $\mathbb{Q}$ with Galois group isomorphic to G.

Hilbert's Theorem gives a sufficient condition for the specialized extension not to collapse. In general, the Galois group of the specialized extension is a subgroup of G (cf. Proposition 19) and may be a proper subgroup of G. It is also known that the fixed field K need not always be a purely transcendental extension of $\mathbb{Q}$. An example of this occurs when G is the cyclic group of order 47.

This theorem can be used to give another proof of Proposition 42:

Corollary. S_n is a Galois group over $\mathbb{Q}$, for all n.

Proof of the corollary. We have already proved that the fixed field of S_n acting in the obvious fashion on $\mathbb{Q}(x_1, \ldots, x_n)$ is purely transcendental over $\mathbb{Q}$ (with the elementary symmetric functions as a transcendence base), so Hilbert's Theorem immediately implies the corollary.

The hypothesis that K be purely transcendental over $\mathbb{Q}$ is crucial to the proof of Hilbert's Theorem. Every finite group is isomorphic to a subgroup of S_n and so acts

on $\mathbb{Q}(x_1, \ldots, x_n)$ for some n. It is not known, however, even for the subgroup A_n of S_n whether its fixed field under the obvious action is a purely transcendental extension of $\mathbb{Q}$ (although it is known by other means that A_n is a Galois group over $\mathbb{Q}$ for all n). Thus there are a number of important open problems in this area of research.

One should also notice that Hilbert's Theorem does not work when the base field $\mathbb{Q}$ is replaced by an arbitrary field F (suppose F were algebraically closed, for instance). In particular, as noted earlier, the general polynomial $f(x)$ in Section 6 has Galois group S_n over $F(a_1, \ldots, a_n)$ for any F, but when F is a finite field, the specialized extension obtained from its splitting field is always cyclic.

We next expand on the theory of inseparable extensions described in Section 13.5. Let p be a prime and let F be a field of characteristic p.

Definition. An algebraic extension E/F is called *purely inseparable* if for each $\alpha \in E$ the minimal polynomial of α over F has only one distinct root.

It is easy to see that the following are equivalent:

(1) E/F is purely inseparable
(2) if $\alpha \in E$ is separable over F, then $\alpha \in F$
(3) if $\alpha \in E$, then $\alpha^{p^n} \in F$ for some n (depending on α), and $m_{\alpha,F}(x) = x^{p^n} - \alpha^{p^n}$.

The following easy proposition describes composites of separable and purely inseparable extensions:

Proposition. If E_1 and E_2 are subfields of E which are both separable (or both purely inseparable) extensions of F, then their composite E_1E_2 is separable (purely inseparable, respectively) over F.

Proof: Exercise.

One immediate consequence of this is:

Proposition. Let E/F be an algebraic extension. Then there is a unique field E_{sep} with $F \subseteq E_{sep} \subseteq E$ such that E_{sep} is separable over F and E is purely inseparable over E_{sep}. The field E_{sep} is the set of elements of E which are separable over F.

The degree of E_{sep}/F is called the *separable degree* of E/F and the degree of E/E_{sep} is called the *inseparable degree* of E/F (often denoted as $[E : F]_s$ and $[E : F]_i$ respectively). The product of these two degrees is the (ordinary) degree. The propositions immediately give:

Corollary. Separable degrees (respectively inseparable degrees) are multiplicative.

When E is generated over F by the root of an irreducible polynomial $p(x) \in F[x]$ the separable and inseparable degrees of the extension E/F are the same as the separable and inseparable degrees of the polynomial $p(x)$ defined in Section 13.5.

The proposition asserts that any algebraic extension may be decomposed into a separable extension followed by a purely inseparable one. An exercise at the end of this section outlines an example illustrating that this decomposition cannot generally be reversed, namely an extension which is not a separable extension of a purely inseparable extension. We shall shortly state conditions on an extension under which the decomposition into separable and purely inseparable subextensions may be reversed.

We now know that an arbitrary extension E/F can be decomposed into a purely transcendental extension $F(S)$ of F followed by a separable extension E_1 of $F(S)$ followed by a purely inseparable extension E/E_1. In certain instances the inseparability in the algebraic extension at the "top" may be removed by a judicious choice of transcendence base:

Proposition. If E is a finitely generated extension of a perfect field F, then there is a transcendence base T of E/F such that E is a separable (algebraic) extension of $F(T)$.

A transcendence base T as described in the proposition is called a *separating transcendence base.* An exercise at the end of this section illustrates this with a nontrivial example.

Recall that an extension E/F is *normal* if it is the splitting field of some (possibly infinite) set of polynomials in $F[x]$ (in particular, normal extensions are algebraic but not necessarily finite or separable). We previously used the synonymous term splitting field and the term normal is reintroduced here in the context of arbitrary algebraic extensions since it is used frequently in the literature, often in the context of embeddings of a field into an algebraic closure. Although the following set of equivalences can be gleaned from the preceding sections, the reader should write out a complete proof, checking that the arguments work for both infinite and inseparable extensions:

Proposition. Let E/F be an arbitrary algebraic extension and let Ω be an algebraic closure of E. The following are equivalent:

(1) E/F is a normal extension (i.e. is the splitting field over F of some set of polynomials in $F[x]$)

(2) whenever $\sigma : E \to \Omega$ is an embedding such that $\sigma|_F$ is the identity, $\sigma(E) = E$

(3) whenever an irreducible polynomial $f(x) \in F[x]$ has one root in E, it has all its roots in E.

In general, any embedding of a normal extension E/F into an algebraic closure of E which extends the identity embedding of F is an automorphism of E, i.e. is an element of $\mathrm{Aut}(E/F)$. Moreover, the number of such automorphisms equals the separable degree of E/F, provided the latter is finite:

$$\text{if } E/F \text{ is a normal extension and } [E:F]_s \text{ is finite,} \quad |\mathrm{Aut}(E/F)| = [E:F]_s.$$

If $[E:F]_s$ is infinite we shall see shortly that $|\mathrm{Aut}(E/F)|$ is also infinite but need not be of the same cardinality.

If E/F is a normal extension whose separable degree is finite, let E_0 be the fixed field of $\mathrm{Aut}(E/F)$. By Corollary 11, E/E_0 is a (separable) Galois extension

whose degree equals $|\mathrm{Aut}(E/F)|$. It follows that E_0/F must be purely inseparable (of degree equal to $[E:F]_i$), i.e. the separable and purely inseparable pieces of the extension may be reversed for normal extensions. More precisely, we easily obtain:

Proposition. If E/F is normal with $[E:F]_s < \infty$, then $E = E_{sep}E_{pi}$, where E_{pi} is a purely inseparable extension of F (E_{pi} consists of all purely inseparable elements of E over F) and $E_{sep} \cap E_{pi} = F$.

Finally, we mention how Galois Theory generalizes to infinite extensions.

Definition. An extension E/F is called *Galois* if it is algebraic, normal and separable. In this case $\mathrm{Aut}(E/F)$ is called the *Galois group* of the extension and is denoted by $\mathrm{Gal}(E/F)$.

For infinite extensions there need not be a bijection between the set of all subgroups of the Galois group and the set of all subfields of E containing F, as the following example illustrates.

Let E be the subfield of $\mathbb{R}$ obtained by adjoining to $\mathbb{Q}$ all square roots of positive rational numbers. One easily sees that E may also be described as the splitting field of the set of polynomials $x^2 - p$, where p runs over all primes in $\mathbb{Z}^+$. Note that E is a (countably) infinite Galois extension of $\mathbb{Q}$. Since every automorphism σ of E is determined by its action on the square roots of the primes and σ either fixes or negates each of these, σ^2 is the identity automorphism. It follows that $\mathrm{Aut}(E)$ is an infinite elementary abelian 2-group. Thus $\mathrm{Aut}(E)$ is an infinite dimensional vector space over $\mathbb{F}_2$. By an exercise in the section on dual spaces (Section 11.3) the number of nonzero homomorphisms of $\mathrm{Aut}(E)$ into $\mathbb{F}_2$ is uncountable, whence their kernels (which are subspaces of co-dimension 1) are uncountable in number (and distinct). Thus $\mathrm{Aut}(E)$ has *uncountably* many subgroups of index 2, whereas $\mathbb{Q}$ has only a *countable* number of quadratic extensions.

The basic problem is that many (most) subgroups of $\mathrm{Gal}(E/F)$ do not correspond (in a bijective fashion) to subfields of E containing F. In order to pick out the "right" set of subgroups of $\mathrm{Gal}(E/F)$ we must introduce a topology on this group (called the Krull topology). The axioms for the collection of (topologically) closed subsets of a topological space are precisely the bookkeeping devices which single out the relevant subgroups. Galois theory for finite extensions force certain subgroups of finite index to be closed sets and these in turn determine the topology on the entire group (as we might expect since every extension of F inside E is a composite of finite extensions).

Theorem. (Krull) Let E/F be a Galois extension with Galois group G. Topologize G by taking as a base for the closed sets the subgroups of G which are the fixing subgroups of the finite extensions of F in E. Then with this ("Krull") topology the closed subgroups of G correspond bijectively with the subfields of E containing F and the correponding lattices are dual. Closed normal subgroups of G correspond to normal extensions of F in E.

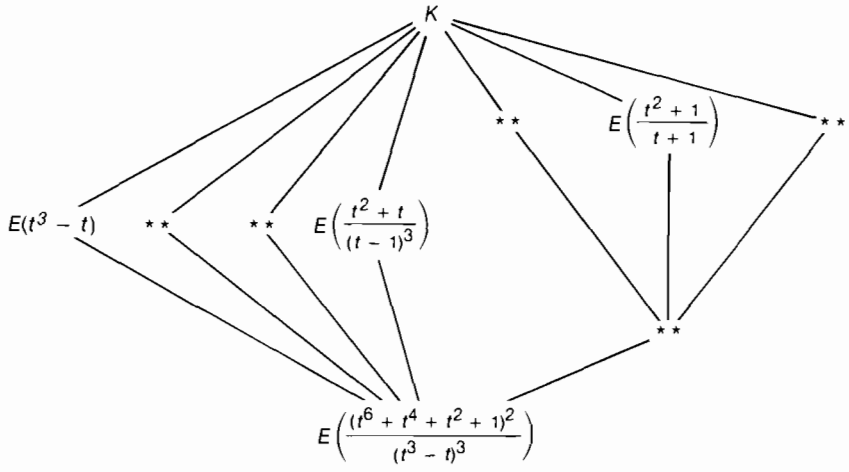
One important area of current research is to describe (as a topological group) the Galois group of certain field extensions such as $\overline{F}/F$, where $\overline{F}$ is the algebraic closure of F. Little is known about the latter group when $F = \mathbb{Q}$ (in particular, its normal subgroups of finite index, i.e. which finite groups occur as Galois groups over $\mathbb{Q}$, are not known). If E is the algebraic closure of the finite field $\mathbb{F}_p$, the Galois group of this extension is the topologically cyclic group $\widehat{\mathbb{Z}}$ with the Frobenius automorphism as a topological generator. The group $\widehat{\mathbb{Z}}$ is an uncountable group (in particular, is not isomorphic to $\mathbb{Z}$) with the property that every closed subgroup of finite index is normal with cyclic quotient. Note that $\widehat{\mathbb{Z}}$ must also have nontrivial infinite closed subgroups (unlike $\mathbb{Z}$) since E contains proper subfields which are infinite over $\mathbb{F}_p$ (such as the composite of all extensions of $\mathbb{F}_p$ of 2-power degree).

EXERCISES

1. Prove that every purely inseparable extension is normal.
2. Let p be a prime and let $K = \mathbb{F}_p(x, y)$ with x and y independent transcendentals over $\mathbb{F}_p$. Let $F = \mathbb{F}_p(x^p - x, y^p - x)$.
   (a) Prove that $[K : F] = p^2$ and the separable degree and inseparable degree of K/F are both equal to p.
   (b) Prove that there is a subfield E of K containing F which is purely inseparable over F of degree p (so then K is a separable extension of E of degree p). [Let $s = x^p - x \in F$ and $t = y^p - x \in F$ and consider $s - t$.]
3. Let p be an odd prime, let s and t be independent transcendentals over $\mathbb{F}_p$, and let $F = \mathbb{F}_p(s, t)$. Let β be a root of $x^2 - sx + t = 0$ and let α be a root of $x^p - \beta = 0$ (in some algebraic closure of F). Set $E = F(\beta)$ and $K = F(\alpha)$.
   (a) Prove that E is a Galois extension of F of degree 2 and that K is a purely inseparable extension of E of degree p.
   (b) Prove that K is not a normal extension of F. [If it were, conjugate β over F to show that K would contain a pth root of s and then also a pth root of t, so $[K : F] \geq p^2$, a contradiction.]
   (c) Prove that there is no field K_0 such that $F \subseteq K_0 \subseteq K$ with K_0/F purely inseparable and K/K_0 separable. [If there were such a field, use Exercise 1 and the fact that the composite of two normal extensions is again normal to show that K would be a normal extension of F.]
4. Under the notation of the previous exercise prove that α, s is a separating transcendence base for K over $\mathbb{F}_p$.
5. Let p be a prime, let t be transcendental over $\mathbb{F}_p$ and let K be obtained by adjoining to $\mathbb{F}_p(t)$ all p-power roots of t. Prove that K has transcendence degree 1 over $\mathbb{F}_p$ and has no separating transcendence base.
6. Show that if t is transcendental over $\mathbb{Q}$, $\mathbb{Q}(t, \sqrt{t^3 - t})$ is not a purely transcendental extension of $\mathbb{Q}$. (This is an example of what is called an *elliptic* function field.)
7. Let k be the field with 4 elements, t a transcendental over k, $F = k(t^4 + t)$ and $K = k(t)$.
   (a) Show that $[K : F] = 4$.
   (b) Show that K is separable over F.
   (c) Show that K is Galois over F.
   (d) Describe the lattice of subgroups of the Galois group and the corresponding lattice of subfields of K, giving each subfield in the form $k(r)$, for some rational function r.

8. Let p be an odd prime, k an algebraically closed field of characteristic p and let t be transcendental over k. Suppose F is a degree 2 field extension of $k(t)$. Show that $F = k(t, y)$, for some $y \in F$ with $y^2 \in k(t)$ and y transcendental over k. If $y^2 = 4t^3 - t - 1$, find $[F : k(y)]$ and describe $k(t) \cap k(y)$ as $k(r)$, for some $r \in k(t)$.

9. Let t be transcendental over $\mathbb{F}_3$, let $K = \mathbb{F}_3(t)$, let $G = \mathrm{Aut}(K/\mathbb{F}_3)$ and let F be the fixed field of G.
   (a) Prove $G \cong S_4$ and deduce that there is a unique field E with $F \subseteq E \subseteq K$ and $[E : F] = 2$. [Recall that $G \cong PGL_2(\mathbb{F}_3)$; show that $GL_2(\mathbb{F}_3)$ permutes the 4 lines in a 2-dimensional vector space over $\mathbb{F}_3$ and the kernel of this permutation representation is the scalar matrices.]
   (b) Describe the lattice of subfields of K containing E, giving each subfield in the form $E(r)$, for some rational function r. (The lattice of subgroups of A_4 appears in Section 2.5). A portion of this lattice is the following:



10. Prove that a purely transcendental proper extension of a field is never algebraically closed.

11. Let S be a set of independent transcendentals over a field F and let Ω be an algebraic closure of $F(S)$. Prove that any permutation on S extends to an element of $\mathrm{Aut}(F(S)/F)$. Prove that any such automorphism of $F(S)$ extends to an automorphism of Ω. Deduce that $\mathbb{C}$ has infinitely many automorphisms.

12. Let K be a subfield of $\mathbb{C}$ maximal with respect to the property: $\sqrt{2} \notin K$.
   (a) Show such a field K exists.
   (b) Show that $\mathbb{C}$ is algebraic over K.
   (c) Prove that every finite extension of K in $\mathbb{C}$ is Galois with Galois group a cyclic 2-group.
   (d) Deduce that $[\mathbb{C} : K]$ is countable (and not finite).

13. Let K be the fixed field in $\mathbb{C}$ of an automorphism of $\mathbb{C}$. Prove that every finite extension of K in $\mathbb{C}$ is cyclic.

14. Let K_n be the splitting field of $(x^2 - p_1)(x^2 - p_2) \cdots (x^2 - p_n)$ over $\mathbb{Q}$, where $p_1, \ldots, p_n$ are the first n primes. Prove that the Galois group of $K_n/\mathbb{Q}$ is an elementary abelian 2-group of order 2^n.

15. Let $K_0 = \mathbb{Q}$ and for $n \geq 0$ define the field K_{n+1} as the extension of K_n obtained by adjoining to K_n all roots of all cubic polynomials over K_n. Let K be the union of the subfields

K_n, $n \geq 0$. Prove that K is a Galois extension of $\mathbb{Q}$. Prove that every cubic polynomial over K splits completely over K. Prove that there are nontrivial algebraic extensions of K.

16. Let F be the composite of all the splitting fields of irreducible cubics over $\mathbb{Q}$. Prove that F does not contain all quadratic extensions of $\mathbb{Q}$.

17. Let $K_0 = \mathbb{Q}$ and for $n \geq 0$ define the field K_{n+1} as the extension of K_n obtained by adjoining to K_n all radicals of elements in K_n. Let K be the union of the subfields K_n, $n \geq 0$. Prove that K is a Galois extension of $\mathbb{Q}$. Prove that there are no nontrivial solvable Galois extensions of K. Prove that there are nontrivial Galois extensions of K.

18. Let $F_0 = \mathbb{Q}$ and for $n \geq 0$ define the field F_{n+1} as the extension of F_n obtained by adjoining to F_n all real radicals of elements in F_n. Let F be the union of the subfields F_n, $n \geq 0$. Let K^+ be the fixed field of complex conjugation restricted to the field K in the previous exercise (the maximal real subfield of K). Prove that $F \neq K^+$.

Part V

INTRODUCTION TO THE REPRESENTATION THEORY OF FINITE GROUPS

The final two chapters are an introduction to the representation theory of finite groups together with some applications. We have already seen in Part I how actions of groups on sets, namely permutation representations, are a fundamental tool for unravelling the structure of groups. Cayley's Theorem and Sylow's Theorem as well as many of the results and applications in Sections 6.1 and 6.2 are based on groups acting on sets (often on subsets of themselves). The chapter on Galois Theory developed one of the most beautiful correspondences in mathematics where the action of a group as automorphisms of a field gives rise to a correspondence between the lattice of subgroups of the Galois group and the lattice of subfields of a Galois extension of fields. In that chapter the interest was more in the field being acted upon than on the acting group. In these final two chapters we study groups acting as linear transformations on vector spaces. Once again we shall be primarily interested in utilizing these linear actions to provide information about the groups themselves. As with the cycle decompositions and signs (odd or even) attached to group elements through permutation representations, we shall find numerical invariants (such as the trace and the determinant) attached to group elements through linear representations. We shall use these invariants to derive group-theoretic information — for example, the existence of normal subgroups. We shall also mention how this theory sits within the encompassing setting of harmonic analysis (the foundations of the theory of Fourier series is a special case) and how linear actions provide an invaluable tool for analyzing vector spaces (especially spaces of functions) which admit a group action. In the theory of group representations the acting group and the space admitting the action are objects of intertwined and comparable interest.

We saw as early as Section 1.7 that there is a direct correspondence between actions of a group on a set and homomorphisms of the group into a symmetric group (these were the permutation representations). More precisely, these are equivalent notions: given an action of a group G on a set A, one obtains a homomorphism $G \to S_A$, and vice versa. A *linear representation* of a group G is any homomorphism of G into $GL(V)$, where V is any vector space over a field F and $GL(V)$ is the group of all nonsingular linear transformations from V to V. Thus a linear representation is a group action in which the set being acted upon has the additional structure of a vector space and the elements of the acting group preserve this structure (i.e. act as linear transformations).

In Part III we saw that modules are the "representation objects" for rings in the sense that the axioms for an R-module specify a "ring action" of R on some abelian group M which preserves the abelian group structure of M. In the case where M was an $F[x]$-module, x acted as a linear transformation from the vector space M to itself. In Chapter 12 the classification of finitely generated modules over Principal Ideal Domains gave us a great deal of information about these linear transformations of M (e.g. canonical forms). In this part we follow a process similar to the study of $F[x]$-modules, replacing the polynomial ring with the group ring FG of G, and we show that there is a correspondence between linear representations of G and FG-modules. We then classify all finitely generated FG-modules and use this classification to describe the linear representations of G. This is analogous to having a "canonical form theorem" for the entire group G of linear transformations (rather than for a single linear transformation, as in the case of the theory of modules over $F[x]$).

There is a natural progression in the three classifications of finitely generated modules over the rings F, $F[x]$ and FG, (i.e. over fields, Principal Ideal Domains and group rings). In each instance the structure and complexity of the modules over that ring reflects the complexity of the ideal structure of the ring. When the ring is a field (which has only the trivial ideals) every module is a direct sum of copies of the field itself. When the ring is a Principal Ideal Domain the ideal structure is, in some sense, the next most trivial — all ideals are cyclic modules. The classification reflects this: finitely generated modules are direct sums of cyclic modules. In this chapter we follow the same procedure of first determining the ideal structure of the group ring FG, then showing how the structure of modules over this ring intimately reflects the ideal structure of the ring itself. This classification is then a natural culmination of a sequence of three related classifications and as such is a fitting topic to conclude our introduction to the main themes in abstract algebra.

15

Representation Theory and Character Theory

15.1 LINEAR ACTIONS AND MODULES OVER GROUP RINGS

For the remainder of the book the groups we consider will be finite groups, unless explicitly mentioned otherwise. Throughout this section F is a field and G is a finite group. We first introduce the basic terminology. Recall that if V is a vector space over F, then $GL(V)$ is the group of nonsingular linear transformations from V to itself (under composition), and if $n \in \mathbb{Z}^+$, then $GL_n(F)$ is the group of invertible $n \times n$ matrices with entries from F (under matrix multiplication).

Definition. Let G be a finite group, let F be a field and let V be a vector space over F.

(1) A *linear representation* of G is any homomorphism from G into $GL(V)$. The *degree* of the representation is the dimension of V.

(2) Let $n \in \mathbb{Z}^+$. A *matrix representation* of G is any homomorphism from G into $GL_n(F)$.

(3) A linear or matrix representation is *faithful* if it is injective.

(4) The *group ring* of G over F is the set of all formal sums of the form

$$\sum_{g \in G} \alpha_g g, \qquad \alpha_g \in F$$

with addition defined componentwise and multiplication defined by $(\alpha g)(\beta h) = (\alpha\beta)(gh)$ (where α and β are multiplied in F and gh is the product in G), and multiplication is extended to sums via the distributive law (cf. Section 7.2).

Unless we are specifically discussing permutation representations the term "representation" will always mean "linear representation." When we wish to emphasize the field F we shall say F-representation, or representation of G on V over F.

Recall that if V is a finite dimensional vector space of dimension n, then by fixing a basis of V we obtain an isomorphism $GL(V) \cong GL_n(F)$. In this way any linear representation of G on a finite dimensional vector space gives a matrix representation and vice versa. For the most part our linear representations will be of finite degree and we shall pass freely between linear representations and matrix representations (specifying a basis when we wish to give an explicit correspondence between the two). Further-

more, given a linear representation $\varphi : G \to GL(V)$ of finite degree, a corresponding matrix representation provides numerical invariants (such as the determinant of $\varphi(g)$, $g \in G$) which are independent of the choice of basis giving the isomorphism between $GL(V)$ and $GL_n(F)$. The exploitation of such invariants will be fundamental to our development.

Before giving examples of representations we recall the group ring FG in greater detail (group rings were introduced in Section 7.2, and some notation and examples were discussed in that section). Suppose the elements of G are $g_1, g_2, \ldots, g_n$. Each element of FG is of the form

$$\sum_{i=1}^{n} \alpha_i g_i, \qquad \alpha_i \in F.$$

Two formal sums[1] are equal if and only if all corresponding coefficients of group elements are equal. Addition and multiplication in FG are defined as follows:

$$\sum_{i=1}^{n} \alpha_i g_i \quad + \quad \sum_{i=1}^{n} \beta_i g_i = \sum_{i=1}^{n} (\alpha_i + \beta_i) g_i$$

$$\left(\sum_{i=1}^{n} \alpha_i g_i\right)\left(\sum_{i=1}^{n} \beta_i g_i\right) = \sum_{k=1}^{n} \left(\sum_{\substack{i,j \\ g_i g_j = g_k}} \alpha_i \beta_j \right) g_k$$

where addition and multiplication of the coefficients α_i and β_j is performed in F. Note that by definition of multiplication,

FG is a commutative ring if and only if G is an abelian group.

The group G appears in FG (identifying g_i with $1g_i$) and the field F appears in FG (identifying β with βg_1, where g_1 is the identity of G). Under these identifications

$$\beta\left(\sum_{i=1}^{n} \alpha_i g_i\right) = \sum_{i=1}^{n} (\beta \alpha_i) g_i, \qquad \text{for all } \beta \in F.$$

In this way

FG is a vector space over F with the elements of G as a basis.

In particular, FG is a vector space over F of dimension equal to $|G|$. The elements of F commute with all elements of FG, i.e. F is in the *center* of FG. When we wish to emphasize the latter two properties we shall say that FG is an *F-algebra* (in general, an F-algebra is a ring R which contains F in its center, so R is both a ring and an F-vector space).

Note that the operations in FG are similar to those in the F-algebra $F[x]$ (although $F[x]$ is infinite dimensional over F). In some works FG is denoted by $F[G]$, although the latter notation is currently less prevalent.

[1] The formal sum displayed above is a way of writing the function from G to F which takes the value α_i on the group element g_i. This same "formality" was used in the construction of free modules (see Theorem 6 in Section 10.3).

Examples

(1) If $G = \langle g \rangle$ is cyclic of order $n \in \mathbb{Z}^+$, then the elements of FG are of the form

$$\sum_{i=0}^{n-1} \alpha_i g^i.$$

The map $F[x] \to F\langle g \rangle$ which sends x^k to g^k for all $k \geq 0$ extends by F-linearity to a surjective ring homomorphism with kernel equal to the ideal generated by $x^n - 1$. Thus

$$F\langle g \rangle \cong F[x]/(x^n - 1).$$

This is an isomorphism of F-algebras, i.e. is a ring isomorphism which is F-linear.

(2) Under the notation of the preceding example let $r = 1 + g + g^2 + \cdots + g^{n-1}$, so r is a nonzero element of $F\langle g \rangle$. Note that $rg = g + g^2 + \cdots + g^{n-1} + 1 = r$, hence $r(1 - g) = 0$. Thus the ring $F\langle g \rangle$ contains zero divisors (provided $n > 1$). More generally, if G is any group of order > 1, then for any nonidentity element $g \in G$, $F\langle g \rangle$ is a subring of FG, so FG also contains zero divisors.

(3) Let $G = S_3$ and $F = \mathbb{Q}$. The elements $r = 5(1\,2) - 7(1\,2\,3)$ and $s = -4(1\,2\,3) + 12(1\,3\,2)$ are typical members of $\mathbb{Q}\,S_3$. Their sum and product are seen to be

$$r + s = 5(1\ 2) - 11(1\ 2\ 3) + 12(1\ 3\ 2)$$
$$rs = -20(2\ 3) + 28(1\ 3\ 2) + 60(1\ 3) - 84$$

(recall that products (compositions) of permutations are computed from right to left). An explicit example of a sum and product of two elements in the group ring $\mathbb{Q}\,D_8$ appears in Section 7.2.

Before giving specific examples of representations we discuss the correspondence between representations of G and FG-modules (after which we can simultaneously give examples of both). This discussion closely parallels the treatment of $F[x]$-modules in Section 10.1.

Suppose first that $\varphi : G \to GL(V)$ is a representation of G on the vector space V over F. As above, write $G = \{g_1, \ldots, g_n\}$, so for each $i \in \{1, \ldots, n\}$, $\varphi(g_i)$ is a linear transformation from V to itself. Make V into an FG-module by defining the action of a ring element on an element of V as follows:

$$\left(\sum_{i=1}^{n} \alpha_i g_i\right) \cdot v = \sum_{i=1}^{n} \alpha_i \varphi(g_i)(v), \qquad \text{for all } \sum_{i=1}^{n} \alpha_i g_i \in FG,\ v \in V.$$

We verify a special case of axiom 2(b) of a module (see Section 10.1) which shows precisely where the fact that φ is a group homomorphism is needed:

$$\begin{aligned}
(g_i g_j) \cdot v &= \varphi(g_i g_j)(v) && \text{(by definition of the action)}\\
&= (\varphi(g_i) \circ \varphi(g_j))(v) && \text{(since } \varphi \text{ is a group homomorphism)}\\
&= \varphi(g_i)(\varphi(g_j)(v)) && \text{(by definition of a composition of linear transformations)}\\
&= g_i \cdot (g_j \cdot v) && \text{(by definition of the action).}
\end{aligned}$$

This argument extends by linearity to arbitrary elements of FG to prove that axiom 2(b) of a module holds in general. It is an exercise to check that the remaining module axioms hold.

Note that F is a subring of FG and the action of the field element α on a vector is the same as the action of the ring element $\alpha 1$ on a vector i.e. the FG-module action extends the F action on V.

Suppose now that conversely we are given an FG-module V. We obtain an associated vector space over F and representation of G as follows. Since V is an FG-module, it is an F-module, i.e. it is a vector space over F. Also, for each $g \in G$ we obtain a map from V to V, denoted by $\varphi(g)$, defined by

$$\varphi(g)(v) = g \cdot v \qquad \text{for all } v \in V,$$

where $g \cdot v$ is the given action of the ring element g on the element v of V. Since the elements of F commute with each $g \in G$ it follows by the axioms for a module that for all $v, w \in V$ and all $\alpha, \beta \in F$ we have

$$\begin{aligned}\varphi(g)(\alpha v + \beta w) &= g \cdot (\alpha v + \beta w)\\ &= g \cdot (\alpha v) + g \cdot (\beta w)\\ &= \alpha(g \cdot v) + \beta(g \cdot w)\\ &= \alpha\varphi(g) + \beta\varphi(w),\end{aligned}$$

that is, for each $g \in G$, $\varphi(g)$ is a linear transformation. Furthermore, it follows by axiom 2(b) of a module that

$$\varphi(g_i g_j)(v) = (\varphi(g_i) \circ \varphi(g_j))(v)$$

(this is essentially the calculation above with the steps reversed). This proves that φ is a group homomorphism (in particular, $\varphi(g^{-1}) = \varphi(g)^{-1}$, so every element of G maps to a nonsingular linear transformation, i.e. $\varphi : G \to GL(V)$).

This discussion has established that there is a bijection between FG-modules and pairs V, φ:

$$\left\{ V \text{ an } FG\text{-module} \right\} \quad \longleftrightarrow \quad \left\{ \begin{array}{c} V \text{ a vector space over } F \\ \text{and} \\ \varphi : G \to GL(V) \text{ a representation} \end{array} \right\}.$$

Giving a representation $\varphi : G \to GL(V)$ on a vector space V over F is therefore equivalent to giving an FG-module V. Under this correspondence we shall say that the module V *affords* the representation φ of G.

Recall from Section 10.1 that if a vector space M is made into an $F[x]$-module via the linear transformation T, then the $F[x]$-submodules of M are precisely the T-stable subspaces of M. In the current situation if V is an FG-module affording the representation φ, then a subspace U of V is called *G-invariant* or *G-stable* if $g \cdot u \in U$ for all $g \in G$ and all $u \in U$ (i.e. if $\varphi(g)(u) \in U$ for all $g \in G$ and all $u \in U$). It follows easily that

the FG-submodules of V are precisely the G-stable subspaces of V.

Examples

(1) Let V be a 1-dimensional vector space over F and make V into an FG-module by letting $gv = v$ for all $g \in G$ and $v \in V$. This module affords the representation $\varphi : G \to GL(V)$ defined by $\varphi(g) = I =$ the identity linear transformation, for all $g \in G$. The corresponding matrix representation (with respect to any basis of V) is the homomorphism of G into $GL_1(F)$ which sends every group element to the 1×1 identity matrix. We shall henceforth refer to this as the *trivial representation* of G. The trivial representation has degree 1 and if $|G| > 1$, it is not faithful.

(2) Let $V = FG$ and consider this ring as a left module over itself. Then V affords a representation of G of degree equal to $|G|$. If we take the elements of G as a basis of V, then each $g \in G$ permutes these basis elements under the left regular permutation representation:

$$g \cdot g_i = gg_i.$$

With respect to this basis of V the matrix of the group element g has a 1 in row i and column j if $gg_j = g_i$, and has 0's in all other positions. This (linear or matrix) representation is called the *regular representation* of G. Note that each nonidentity element of G induces a nonidentity permutation on the basis of V so the regular representation is always faithful.

(3) Let $n \in \mathbb{Z}^+$, let $G = S_n$ and let V be an n-dimensional vector space over F with basis $e_1, e_2, \ldots, e_n$. Let S_n act on V by defining for each $\sigma \in S_n$

$$\sigma \cdot e_i = e_{\sigma(i)}, \qquad 1 \le i \le n$$

i.e. σ acts by permuting the subscripts of the basis elements. This provides an (injective) homomorphism of S_n into $GL(V)$ (i.e. a faithful representation of S_n of degree n), hence makes V into an FS_n-module. As in the preceding example, the matrix of σ with respect to the basis $e_1, \ldots, e_n$ has a 1 in row i and column j if $\sigma \cdot e_j = e_i$ (and has 0 in all other entries). Thus σ has a 1 in row i and column j if $\sigma(j) = i$.

For an example of the ring action, consider the action of FS_3 on the 3-dimensional vector space over F with basis e_1, e_2, e_3. Let σ be the transposition $(1\,2)$, let τ be the 3-cycle $(1\,2\,3)$ and let $r = 2\sigma - 3\tau \in FS_3$. Then

$$\begin{aligned} r \cdot (\alpha e_1 + \beta e_2 + \gamma e_3) &= 2(\alpha e_{\sigma(1)} + \beta e_{\sigma(2)} + \gamma e_{\sigma(3)}) - 3(\alpha e_{\tau(1)} + \beta e_{\tau(2)} + \gamma e_{\tau(3)}) \\ &= 2(\alpha e_2 + \beta e_1 + \gamma e_3) - 3(\alpha e_2 + \beta e_3 + \gamma e_1) \\ &= (2\beta - 3\gamma)e_1 - \alpha e_2 + (2\gamma - 3\beta)e_3. \end{aligned}$$

(4) If $\psi : H \to GL(V)$ is any representation of H and $\varphi : G \to H$ is any group homomorphism, then the composition $\psi \circ \varphi$ is a representation of G. For example, let V be the FS_n-module of dimension n described in the preceding example. If $\pi : G \to S_n$ is any permutation representation of G, the composition of π with the representation above gives a linear representation of G. In other words, V becomes an FG-module under the action

$$g \cdot e_i = e_{\pi(g)(i)}, \qquad \text{for all } g \in G.$$

Note that the regular representation, (2), is just the special case of this where $n = |G|$ and π is the left regular permutation representation of G.

(5) Any homomorphism of G into the multiplicative group $F^\times = GL_1(F)$ is a degree 1 (matrix) representation. For example, suppose $G = \langle g \rangle \cong Z_n$ is the cyclic group of order n and ζ is a fixed n^{th} root of 1 in F. Let $g^i \mapsto \zeta^i$, for all $i \in \mathbb{Z}$. This

representation of $\langle g \rangle$ is a faithful representation if and only if ζ is a primitive n^{th} root of 1.

(6) In many situations it is easier to specify an explicit matrix representation of a group G rather than to exhibit an FG-module. For example, recall that the dihedral group D_{2n} has the presentation

$$D_{2n} = \langle r, s \mid r^n = s^2 = 1,\ rs = sr^{-1} \rangle.$$

If R and S are any matrices satisfying the relations $R^n = S^2 = I$ and $RS = SR^{-1}$ then the map $r \mapsto R$ and $s \mapsto S$ extends uniquely to a homomorphism from D_{2n} to the matrix group generated by R and S, hence gives a representation of D_{2n}. An explicit example of matrices $R, S \in M_2(\mathbb{R})$ may be obtained as follows. If a regular n-gon is drawn on the x, y plane centered at the origin with the line $y = x$ as one of its lines of symmetry then the matrix R that rotates the plane through $2\pi/n$ radians and the matrix S that reflects the plane about the line $y = x$ both send this n-gon onto itself. It follows that these matrices act as symmetries of the n-gon and so satisfy the above relations. These matrices are readily computed (cf. Exercise 25, Section 1.6) and so the maps

$$r \mapsto R = \begin{pmatrix} \cos 2\pi/n & -\sin 2\pi/n \\ \sin 2\pi/n & \cos 2\pi/n \end{pmatrix} \qquad \text{and} \qquad s \mapsto S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

extend uniquely to a (degree 2) representation of D_{2n} into $GL_2(\mathbb{R})$. Since the matrices R and S have orders n and 2 respectively, it follows that they generate a subgroup of $GL_2(\mathbb{R})$ of order $2n$ and hence this representation is faithful.

(7) By using the usual generators and relations for the quaternion group

$$Q_8 = \langle i, j \mid i^4 = j^4 = 1,\ i^2 = j^2,\ i^{-1}ji = j^{-1} \rangle$$

one may similarly obtain (cf. Exercise 26, Section 1.6) a representation $\varphi : Q_8 \to GL_2(\mathbb{C})$ defined by

$$\varphi(i) = \begin{pmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{pmatrix} \qquad \text{and} \qquad \varphi(j) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

This representation of Q_8 is faithful.

(8) A 4-dimensional representation of the quaternion group Q_8 may be obtained from the real Hamilton quaternions, $\mathbb{H}$ (cf. Section 7.1). The group Q_8 is a subgroup of the multiplicative group of units of $\mathbb{H}$ and each of the elements of Q_8 acts by left multiplication on the 4-dimensional real vector space $\mathbb{H}$. Since the real numbers are in the center of $\mathbb{H}$ (i.e. since $\mathbb{H}$ is an $\mathbb{R}$-algebra), left multiplication is $\mathbb{R}$-linear. This linear action thus gives a homomorphism from Q_8 into $GL_4(\mathbb{R})$. One can easily write out the explicit matrices of each of the elements of Q_8 with respect to the basis $1, i, j, k$ of $\mathbb{H}$. For example, left multiplication by i acts by $1 \mapsto i$, $i \mapsto -1$, $j \mapsto k$ and $k \mapsto -j$ and left multiplication by j acts by $1 \mapsto j$, $i \mapsto -k$, $j \mapsto -1$ and $k \mapsto i$ so

$$i \mapsto \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{and} \qquad j \mapsto \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

This representation of Q_8 is also faithful.

(9) Suppose that H is a normal subgroup of the group G and suppose that H is an elementary abelian p-group for some prime p. Then $V = H$ is a vector space over $\mathbb{F}_p$, where the scalar a acts on the vector v by $av = v^a$ (see Section 10.1). The action of each element of G by conjugation on V is $\mathbb{F}_p$-linear because $gv^a g^{-1} = (gvg^{-1})^a$ and this action of G on V makes V into an $\mathbb{F}_p G$-module (the automorphisms of elementary abelian p-groups were discussed in Sections 4.4 and 10.1). The kernel of this representation is the set of elements of G that commute with every element of H, $C_G(H)$ (which always contains the abelian group H itself). Thus the action of a group on subsets of itself often affords linear representations over finite fields. Representations of groups over finite fields are called *modular representations* and these are fundamental to the study of the internal structure of groups.

(10) For an example of an FG-submodule, let $G = S_n$ and let V be the FS_n-module described in Example 3. Let N be the subspace of V consisting of vectors all of whose coordinates are equal, i.e.

$$N = \{\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n \mid \alpha_1 = \alpha_2 = \cdots = \alpha_n\}$$

(this is a 1-dimensional S_n-stable subspace). Each $\sigma \in S_n$ fixes each vector in N so the submodule N affords the trivial representation of S_n. As an exercise, one may show that if $n \geq 3$ then N is the *unique* 1-dimensional subspace of V which is S_n-stable, i.e. N is the unique 1-dimensional FS_n-submodule (N is called the *trace* submodule of FS_n).

Another FS_n-submodule of V is the subspace I of all vectors whose coordinates sum to zero:

$$I = \{\alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n \mid \alpha_1 + \alpha_2 + \cdots + \alpha_n = 0\}.$$

Again I is an S_n-stable subspace (since each $\sigma \in S_n$ permutes the coordinates of each vector in V, each σ leaves the sum of the coefficients unchanged). Since I is the kernel of the linear transformation from V onto F which sends a vector to the sum of its coefficients (called the augmentation map – cf. Section 7.3), I has dimension $n - 1$.

(11) If $V = FG$ is the regular representation of G described in Example 2 above, then V has FG-submodules of dimensions 1 and $|G| - 1$ as in the preceding example:

$$N = \{\alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_n g_n \mid \alpha_1 = \alpha_2 = \cdots = \alpha_n\}$$
$$I = \{\alpha_1 g_1 + \alpha_2 g_2 + \cdots + \alpha_n g_n \mid \alpha_1 + \alpha_2 + \cdots + \alpha_n = 0\}.$$

In fact N and I are 2-sided ideals of FG (not just left ideals — note that N is in the center of FG). The ideal I is called the *augmentation ideal* of FG and N is called the *trace ideal* of FG.

Recall that in the study of a linear transformation T of a vector space V to itself we made V into an $F[x]$-module (where x acted as T on V); our goal was to decompose V into a direct sum of cyclic submodules. In this way we were able to find a basis of V for which the matrix of T with respect to this basis was in some *canonical* form. Changing the basis of V did not change the module V but changed the matrix representation of T by similarity (i.e. changed the isomorphism between $GL(V)$ and $GL_n(F)$). We introduce the analogous terminology to describe when two FG-modules are the same up to a change of basis.

Definition. Two representations of G are *equivalent* (or *similar*) if the FG-modules affording them are isomorphic modules. Representations which are not equivalent are called *inequivalent*.

Suppose $\varphi : G \to GL(V)$ and $\psi : G \to GL(W)$ are equivalent representations (here V and W must be vector spaces over the same field F). Let $T : V \to W$ be an FG-module isomorphism between them. Since T is, in particular, an F-module isomorphism, T is a vector space isomorphism, so V and W must have the same dimension. Furthermore, for all $g \in G$, $v \in V$ we have $T(g \cdot v) = g \cdot (T(v))$, since T is an isomorphism of FG-modules. By definition of the action of ring elements this means $T(\varphi(g)v) = \psi(g)(T(v))$, that is

$$T \circ \varphi(g) = \psi(g) \circ T \qquad \text{for all } g \in G.$$

In particular, if we identify V and W as vector spaces, then

$$\begin{gathered} \textit{two representations } \varphi \textit{ and } \psi \textit{ of } G \textit{ on a vector space } V \textit{ are equivalent} \\ \textit{if and only if there is some } T \in GL(V) \textit{ such that} \\ T \circ \varphi(g) \circ T^{-1} = \psi(g) \qquad \textit{for all } g \in G. \end{gathered}$$

This T is a *simultaneous* change of basis for all $\varphi(g)$, $g \in G$.

In matrix terminology, two representations φ and ψ are equivalent if there is a fixed invertible matrix P such that

$$P\varphi(g)P^{-1} = \psi(g) \qquad \text{for all } g \in G.$$

The linear transformation T or the matrix P above is said to *intertwine* the representations φ and ψ (it gives the "rule" for changing φ into ψ).

In order to study the decomposition of an FG-module into (direct sums of) submodules we shall need some terminology. We state these definitions for arbitrary rings since we shall be discussing direct sum decompositions in greater generality in the next section.

Definition. Let R be a ring and let M be a nonzero R-module.

(1) The module M is said to be *irreducible* (or *simple*) if its only submodules are 0 and M; otherwise M is called *reducible*.

(2) The module M is said to be *indecomposable* if M cannot be written as $M_1 \oplus M_2$ for any nonzero submodules M_1 and M_2; otherwise M is called *decomposable*.

(3) The module M is said to be *completely reducible* if it is a direct sum of irreducible submodules.

(4) A representation is called *irreducible, reducible, indecomposable, decomposable* or *completely reducible* according to whether the FG-module affording it has the corresponding property.

(5) If M is a completely reducible R-module, any direct summand of M is called a *constituent* of M (i.e. N is a constituent of M if there is a submodule N' of M such that $M = N \oplus N'$).

An irreducible module is, by definition, both indecomposable and completely reducible. We shall shortly give examples of indecomposable modules that are not irreducible. The next section explores these notions in greater detail for general rings, giving examples and further explanation of the relations between them; for the moment we interpret these terms in the context of representations.

If $R = FG$, an irreducible FG-module V is a nonzero F-vector space with no nontrivial, proper G-invariant subspaces. For example, if $\dim_F V = 1$ then V is necessarily irreducible (its only subspaces are 0 and V).

Suppose V is a finite dimensional FG-module and V is reducible. Let U be a G-invariant subspace. Form a basis of V by taking a basis of U and enlarging it to a basis of V. Then for each $g \in G$ the matrix, $\varphi(g)$, of g acting on V with respect to this basis is of the form

$$\varphi(g) = \begin{pmatrix} \varphi_1(g) & \psi(g) \\ 0 & \varphi_2(g) \end{pmatrix}$$

where $\varphi_1 = \varphi|_U$ (with respect to the chosen basis of U) and φ_2 is the representation of G on V/U (and ψ is not necessarily a homomorphism — $\psi(g)$ need not be a square matrix). So reducible representations are those with a corresponding matrix representation whose matrices are in block upper triangular form.

Assume further that the FG-module V is decomposable, $V = U \oplus U'$. Take for a basis of V the union of a basis of U and a basis of U'. With this choice of basis the matrix for each $g \in G$ is of the form

$$\varphi(g) = \begin{pmatrix} \varphi_1(g) & 0 \\ 0 & \varphi_2(g) \end{pmatrix}$$

(i.e. $\psi(g) = 0$ for all $g \in G$). Thus decomposable representations are those with a corresponding matrix representation whose matrices are in block diagonal form.

Examples

(1) As noted above, all degree 1 representations are irreducible, indecomposable and completely reducible. In particular, this applies to the trivial representation and to the representations described in Example 5 above.

(2) If $|G| > 1$, the regular representation of G is reducible (the augmentation ideal and the trace ideal are proper nonzero submodules). We shall later determine the conditions under which this representation is completely reducible and how it decomposes into a direct sum.

(3) For $n > 1$ the FS_n-module described in Example 10 above is reducible since N and I are proper, nonzero submodules. The module N is irreducible (being 1-dimensional) and if the characteristic of the field F does not divide n, then I is also irreducible.

(4) The degree 2 representation of the dihedral group $D_{2n} = G$ described in Example 6 above is irreducible for $n \geq 3$. There are no G-invariant 1-dimensional subspaces since a rotation by $2\pi/n$ radians sends no line in $\mathbb{R}^2$ to itself. Similarly, the degree 2 complex representation of Q_8 described in Example 7 is irreducible since the given matrix $\varphi(i)$ has exactly two 1-dimensional eigenspaces (corresponding to its distinct eigenvalues $\pm\sqrt{-1}$) and these are not invariant under the matrix $\varphi(j)$. The degree 4 representation $\varphi : Q_8 \to GL_4(\mathbb{R})$ described in Example 8 can also be shown to be irreducible (see the exercises). We shall see, however, that if we view φ as a complex representation $\varphi : Q_8 \to GL_4(\mathbb{C})$ (just by considering the real entries

of the matrices to be complex entries) then there is a *complex* matrix P such that $P^{-1}\varphi(g)P$ is a direct sum of 2×2 block matrices for all $g \in Q_8$. Thus an irreducible representation over a field F may become reducible when the field is extended.

(5) Let $G = \langle g \rangle$ be cyclic of order n and assume F contains all the nth roots of 1. As noted in Example 1 in the set of examples of group algebras, $F\langle g \rangle \cong F[x]/(x^n - 1)$. Thus the FG-modules are precisely the $F[x]$-modules annihilated by $x^n - 1$. The latter (finite dimensional) modules are described, up to equivalence, by the Jordan Canonical Form Theorem.

If the minimal polynomial of g acting on an $F\langle g \rangle$-module V has distinct roots, there is a basis of V such that g (hence all its powers) is represented by a diagonal matrix (cf. Corollary 25, Section 12.3). In this case, V is a completely reducible $F\langle g \rangle$-module (being a direct sum of 1-dimensional $\langle g \rangle$-invariant subspaces). In general, the minimal polynomial of g acting on V divides $x^n - 1$ so if $x^n - 1$ has distinct roots in F, then V is a completely reducible $F\langle g \rangle$-module. The polynomial $x^n - 1$ has distinct roots in F if and only if the characteristic of F does not divide n. This gives a sufficient condition for every $F\langle g \rangle$-module to be completely reducible.

If the minimal polynomial of g acting on V does *not* have distinct roots (so the characteristic of F does divide n), the Jordan canonical form of g must have an elementary Jordan block of size > 1. Since every linear transformation has a unique Jordan canonical form, g cannot be represented by a diagonal matrix, i.e. V is not completely reducible. It follows from results on cyclic modules in Section 12.3 that the (1-dimensional) eigenspace of g in any Jordan block of size > 1 admits no $\langle g \rangle$-invariant complement, i.e. V is reducible but not completely reducible.

Specifically, let p be a prime, let $F = \mathbb{F}_p$ and let g be of order p. Let V be the 2-dimensional space over $\mathbb{F}_p$ with basis v, w and define an action of g on V by

$$g \cdot v = v \qquad \text{and} \qquad g \cdot w = v + w.$$

This endomorphism of V does have order p (in $GL(V)$) and the matrix of g with respect to this basis is the elementary Jordan block

$$\varphi(g) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Now V is reducible (span$\{v\}$ is a $\langle g \rangle$-invariant subspace) but V is indecomposable (the above 2×2 elementary Jordan matrix is not similar to a diagonal matrix).

The first fundamental result in the representation theory of finite groups shows how Example 5 generalizes to noncyclic groups.

Theorem 1. (Maschke) Let G be a finite group and let F be a field whose characteristic does not divide $|G|$. If V is any FG-module and U is any submodule of V, then V has a submodule W such that $V = U \oplus W$ (i.e. every submodule is a direct summand).

Remark: The hypothesis of Maschke's Theorem applies to any finite group when F has characteristic 0.

Proof: The idea of the proof of Maschke's Theorem is to produce an FG-module homomorphism

$$\pi : V \to U$$

which is a projection onto U, i.e. which satisfies the following two properties:

(i) $\pi(u) = u$ for all $u \in U$
(ii) $\pi(\pi(v)) = \pi(v)$ for all $v \in V$ (i.e. $\pi^2 = \pi$)

(in fact (ii) is implied by (i) and the fact that $\pi(V) \subseteq U$).

Suppose first that we can produce such an FG-module homomorphism and let $W = \ker \pi$. Since π is a module homomorphism, W is a submodule. We see that W is a direct sum complement to U as follows. If $v \in U \cap W$ then by (i), $v = \pi(v)$ whereas by definition of W, $\pi(v) = 0$. This shows $U \cap W = 0$. To show $V = U + W$ let v be an arbitrary element of V and write $v = \pi(v) + (v - \pi(v))$. By definition, $\pi(v) \in U$. By property (ii) of π,

$$\pi(v - \pi(v)) = \pi(v) - \pi(\pi(v)) = \pi(v) - \pi(v) = 0,$$

i.e. $v - \pi(v) \in W$. This shows $V = U + W$ and hence $V = U \oplus W$. To establish Maschke's Theorem it therefore suffices to find such an FG-module projection π.

Since U is a subspace it has a vector space direct sum complement W_0 in V (take a basis $\mathcal{B}_1$ of U, build it up to a basis $\mathcal{B}$ of V and let W_0 be the span of $\mathcal{B} - \mathcal{B}_1$). Thus $V = U \oplus W_0$ as vector spaces but W_0 need not be G-stable (i.e. need not be an FG-submodule). Let $\pi_0 : V \to U$ be the vector space projection of V onto U associated to this direct sum decomposition, i.e. π_0 is defined by

$$\pi_0(u + w) = u \qquad \text{for all } u \in U,\ w \in W_0.$$

The key idea of the proof is to "average" π_0 over G to form an FG-module projection π. For each $g \in G$ define

$$g\pi_0 g^{-1} : V \to U \qquad \text{by} \qquad g\pi_0 g^{-1}(v) = g \cdot \pi_0(g^{-1} \cdot v), \quad \text{for all } v \in V$$

(here $\cdot$ denotes the action of elements of the ring FG). Since π_0 maps V into U and U is stable under the action of g we have that $g\pi_0 g^{-1}$ maps V into U. Both g and g^{-1} act as F-linear transformations, so $g\pi_0 g^{-1}$ is a linear transformation. Furthermore, if u is in the G-stable space U then so is $g^{-1}u$, and by definition of π_0 we have $\pi_0(g^{-1}u) = g^{-1}u$. From this we obtain that for all $g \in G$,

$$g\pi_0 g^{-1}(u) = u \qquad \text{for all } u \in U$$

(i.e. $g\pi_0 g^{-1}$ is also a vector space projection of V onto U).

Let $n = |G|$ and view n as an element of F ($n = 1 + \cdots + 1$, n times). By hypothesis n is not zero in F and so has an inverse in F. Define

$$\pi = \frac{1}{n} \sum_{g \in G} g\pi_0 g^{-1}.$$

Since π is a scalar multiple of a sum of linear transformations from V to U, it is also a linear transformation from V to U. Furthermore, each term in the sum defining π

restricts to the identity map on the subspace U and so $\pi|_U$ is $1/n$ times the sum of n copies of the identity. These observations prove:

$$\begin{aligned} \pi : V \to U &\text{ is a linear transformation} \\ \pi(u) = u &\qquad \text{for all } u \in U \\ \pi^2(v) = \pi(v) &\qquad \text{for all } v \in V. \end{aligned}$$

It remains to show that π is an FG-module homomorphism (i.e. is FG-linear). It suffices to prove that for all $h \in G$, $\pi(hv) = h\pi(v)$, for $v \in V$. In this case

$$\begin{aligned} \pi(hv) &= \frac{1}{n} \sum_{g \in G} g\pi_0(g^{-1}hv) \\ &= \frac{1}{n} \sum_{g \in G} h(h^{-1}g)\pi_0((g^{-1}h)v) \\ &= \frac{1}{n} \sum_{\substack{k = h^{-1}g \\ g \in G}} h(k\pi_0(k^{-1}v) \\ &= h\pi(v) \end{aligned}$$

(as g runs over all elements of G, so does $k = h^{-1}g$ and the module element h may be brought outside the summation by the distributive law in modules). This establishes the existence of the FG-module projection π and so completes the proof.

The applications of Maschke's Theorem will be to finitely generated FG-modules. Unlike the situation of $F[x]$-modules, however, finitely generated FG-modules are automatically finite dimensional vector spaces (the difference being that FG itself is finite dimensional, whereas $F[x]$ is not). Let V be an FG-module. If V is a finite dimensional vector space over F, then a fortiori V is finitely generated as an FG-module (any F basis gives a set of generators over FG). Conversely, if V is finitely generated as an FG-module, say by $v_1, \ldots, v_k$, then one easily sees that V is spanned as a vector space by the finite set $\{g \cdot v_i \mid g \in G,\ 1 \le i \le k\}$. Thus

an FG-module is finitely generated if and only if it is finite dimensional.

Corollary 2. If G is a finite group and F is a field whose characteristic does not divide $|G|$, then every finitely generated FG-module is completely reducible (equivalently, every F-representation of G of finite degree is completely reducible).

Proof: Let V be a finitely generated FG-module. As noted above, V is finite dimensional over F, so we may proceed by induction on its dimension. If V is irreducible, it is completely reducible and the result holds. Suppose therefore that V has a proper, nonzero FG-submodule U. By Maschke's Theorem U has an FG-submodule complement W: $V = U \oplus W$. By induction, each of U and W are direct sums of irreducible submodules, hence so is V. This completes the induction.

Corollary 3. Let G be a finite group, let F be a field whose characteristic does not divide $|G|$ and let $\varphi : G \to GL(V)$ be a representation of G of finite degree. Then

there is a basis of V such that for each $g \in G$ the matrix of $\varphi(g)$ with respect to this basis is block diagonal:

$$\begin{pmatrix} \varphi_1(g) & & & \\ & \varphi_2(g) & & \\ & & \ddots & \\ & & & \varphi_m(g) \end{pmatrix}$$

where φ_i is an irreducible matrix representation of G, $1 \leq i \leq m$.

Proof: By Corollary 2 we may write $V = U_1 \oplus U_2 \oplus \cdots \oplus U_m$, where U_i is an irreducible FG-submodule of V. Let $\mathcal{B}_i$ be a basis of U_i and let $\mathcal{B}$ be the union of the $\mathcal{B}_i$'s. For each $g \in G$, the matrix of $\varphi(g)$ with respect to the basis $\mathcal{B}$ is of the form in the corollary, where $\varphi_i(g)$ is the matrix of $\varphi(g)|_{U_i}$ with respect to the basis $\mathcal{B}_i$.

The converse of Maschke's Theorem is also true. Namely, if the characteristic of F does divide $|G|$, then G possesses (finitely generated) FG-modules which are not completely reducible. Specifically, the regular representation (i.e. the module FG itself) is not completely reducible.

In Section 15.3 we shall discuss the question of uniqueness of the constituents in direct sum decompositions of FG-modules into irreducible submodules.

EXERCISES

Let F be a field, let G be a finite group and let $n \in \mathbb{Z}^+$.

1. Prove that if $\varphi : G \to GL(V)$ is any representation, then φ gives a faithful representation of $G/\ker \varphi$.
2. Let $\varphi : G \to GL_n(F)$ be a matrix representation. Prove that the map $g \mapsto \det(\varphi(g))$ is a degree 1 representation.
3. Prove that the degree 1 representations of G are in bijective correspondence with the degree 1 representations of the abelian group G/G' (where G' is the commutator subgroup of G).
4. Let V be a (possibly infinite dimensional) FG-module (G is a finite group). Prove that for each $v \in V$ there is an FG-submodule containing v of dimension $\leq |G|$.
5. Prove that if $|G| > 1$ then every irreducible FG-module has dimension $< |G|$.
6. Write out the matrices $\varphi(g)$ for every $g \in G$ for each of the following representations that were described in the second set of examples:
   (a) the representation of S_3 described in Example 3 (let $n = 3$ in that example)
   (b) the representation of D_8 described in Example 6 (i.e. let $n = 4$ in that example and write out the values of all the sines and cosines, for all group elements)
   (c) the representation of Q_8 described in Example 7
   (d) the representation of Q_8 described in Example 8.
7. Let V be the 4-dimensional permutation module for S_4 described in Example 3 of the second set of examples. Let $\pi : D_8 \to S_4$ be the permutation representation of D_8 obtained from the action of D_8 by left multiplication on the set of left cosets of its subgroup $\langle s \rangle$. Make V into an FD_8-module via π as described in Example 4 and write out the 4×4 matrices for r and s given by this representation with respect to the basis $e_1, \ldots, e_4$.
8. Let V be the FS_n-module described in Examples 3 and 10 in the second set of examples.
   (a) Prove that if v is any element of V such that $\sigma \cdot v = v$ for all $\sigma \in S_n$ then v is an F-multiple of $e_1 + e_2 + \cdots + e_n$.

(b) Prove that if $n \geq 3$, then V has a unique 1-dimensional submodule, namely the submodule N consisting of all F-multiples of $e_1 + e_2 + \cdots + e_n$.

9. Prove that the 4-dimensional representation of Q_8 on $\mathbb{H}$ described in Example 8 in the second set of examples is irreducible. [Show that any Q_8-stable subspace is a left ideal.]

10. Prove that $GL_2(\mathbb{R})$ has no subgroup isomorphic to Q_8. [This may be done by direct computation using generators and relations for Q_8. Simplify these calculations by putting one generator in rational canonical form.]

11. Let $\varphi : S_n \to GL_n(F)$ be the matrix representation given by the permutation module described in Example 3 in the second set of examples, where the matrices are computed with respect to the basis $e_1, \ldots, e_n$. Prove that $\det \varphi(\sigma) = \epsilon(\sigma)$ for all $\sigma \in S_n$, where $\epsilon(\sigma)$ is the sign of the permutation σ. [Check this on transpositions.]

12. Assume the characteristic of F is not 2. Let H be the set of $T \in M_n(F)$ such that T has exactly one nonzero entry in each row and each column and zeros elsewhere, and the nonzero entries are ± 1. Prove that H is a subgroup of $GL_n(F)$ and that H is isomorphic to $E_{2^n} \rtimes S_n$ (semidirect product), where E_{2^n} is the elementary abelian group of order 2^n.

The next few exercises explore an important result known as Schur's Lemma and some of its consequences.

13. Let R be a ring and let M and N be simple (i.e. irreducible) R-modules.
    (a) Prove that every nonzero R-module homomorphism from M to N is an isomorphism. [Consider its kernel and image.]
    (b) Prove Schur's Lemma: if M is a simple R-module then $\mathrm{Hom}_R(M, M)$ is a division ring (recall that $\mathrm{Hom}_R(M, M)$ is the ring of all R-module homomorphisms from M to M, where multiplication in this ring is composition).

14. Let $\varphi : G \to GL(V)$ be a representation of G. The *centralizer* of φ is defined to be the set of all linear transformations, A, from V to itself such that $A\varphi(g) = \varphi(g)A$ for all $g \in G$ (i.e. the linear transformations of V which commute with all $\varphi(g)$'s).
    (a) Prove that a linear transformation A from V to V is in the centralizer of φ if and only if it is an FG-module homomorphism from V to itself (so the centralizer of φ is the same as the *ring* $\mathrm{Hom}_{FG}(V, V)$).
    (b) Show that if z is in the center of G then $\varphi(z)$ is in the centralizer of φ.
    (c) Assume φ is an irreducible representation (so V is a simple FG-module). Prove that if H is any finite *abelian* subgroup of $GL(V)$ such that $A\varphi(g) = \varphi(g)A$ for all $A \in H$ then H is cyclic (in other words, any finite abelian subgroup of the multiplicative group of units in the ring $\mathrm{Hom}_{FG}(V, V)$ is cyclic). [By the preceding exercise, $\mathrm{Hom}_{FG}(V, V)$ is a division ring, so this reduces to proving that a finite abelian subgroup of the multiplicative group of nonzero elements in a division ring is cyclic. Show that the division subring generated by an abelian subgroup of any division ring is a field and use Proposition 18, Section 9.5.]
    (d) Show that if φ is a faithful irreducible representation then the center of G is cyclic.
    (e) Deduce from (d) that if G is abelian and φ is any irreducible representation then $G/\ker \varphi$ is cyclic.

15. Exhibit all 1-dimensional complex representations of a finite cyclic group; make sure to decide which are inequivalent.

16. Exhibit all 1-dimensional complex representations of a finite abelian group. Deduce that the number of inequivalent degree 1 complex representations of a finite abelian group equals the order of the group. [First decompose the abelian group into a direct product of cyclic groups, then use the preceding exercise.]

17. Prove the following variant of Schur's Lemma for complex representations of abelian groups: if G is abelian, any irreducible complex representation, φ, of G is of degree 1 and $G/\ker\varphi$ is cyclic. [This can be done without recourse to Exercise 14 by using the observation that for any $g \in G$ the eigenspaces of $\varphi(g)$ are G-stable. Your proof should also work for infinite abelian groups.]
18. Prove the general form of Schur's Lemma for complex representations: if $\varphi : G \to GL_n(\mathbb{C})$ is an irreducible matrix representation and A is an $n \times n$ matrix commuting with $\varphi(g)$ for all $g \in G$, then A is a scalar matrix. Deduce that if φ is a faithful, irreducible, complex representation then the center of G is cyclic and $\varphi(z)$ is a scalar matrix for all elements z in the center of G. [As in the preceding exercise, the eigenspaces of A are G-stable.]
19. Prove that if G is an abelian group then any finite dimensional complex representation of G is equivalent to a representation into diagonal matrices (i.e. any finite group of commuting matrices over $\mathbb{C}$ can be simultaneously diagonalized). [This can be done without recourse to Maschke's Theorem by looking at eigenspaces.]
20. Prove that the number of degree 1 complex representations of any finite group G equals $|G : G'|$, where G' is the commutator subgroup of G. [Use Exercises 3 and 16.]
21. Let G be a noncyclic abelian group acting by conjugation on an elementary abelian p-group V, where p is a prime not dividing the order of G.
    (a) Prove that if W is an irreducible $\mathbb{F}_pG$-submodule of V then there is some nonidentity element $g \in G$ such that $W \leq C_V(g)$ (here $C_V(g)$ is the subgroup of elements of V that are fixed by g under conjugation).
    (b) Prove that V is generated by the subgroups $C_V(g)$ as g runs over all nonidentity elements of G.
22. Let p be a prime, let P be a p-group and let F be a field of characteristic p. Prove that the only irreducible representation of P over F is the trivial representation. [Do this for a group of order p first using the fact that F contains all p^{th} roots of 1 (namely 1 itself). If P is not of order p, let z be an element of order p in the center of P, prove that z is in the kernel of the irreducible representation and apply induction to $P/\langle z\rangle$.]
23. Let p be a prime, let P be a nontrivial p-group and let F be a field of characteristic p. Prove that the regular representation is not completely reducible. [Use the preceding exercise.]
24. Let p be a prime, let P be a nontrivial p-group and let F be a field of characteristic p. Prove that the regular representation is indecomposable.

15.2 PROJECTIVE AND INJECTIVE MODULES

This section introduces the notions of projective and injective modules for an arbitrary ring R and gives some examples. These notions pervade the theory of rings and representation theory. Readers wishing to cover basic representation theory as quickly as possible may give this section a cursory reading.

Let R be an arbitrary ring and let A and B be R-modules. One of the basic problems in the study of R-modules is to describe in terms of A and B all ways of "putt ıg A and B together" to form a larger module M in such a way that M has a submodule isomorphic to A whose quotient module is isomorphic to B. This general problem falls under the rubric of *extension theory*, a subject which was touched on for groups in Part I (see the second part of the Hölder Program). One method for constructing such a module M is to take the direct product $A \times B$ (if we identify A and B with their respective component submodules, this is the same as $A \oplus B$). It is

natural to wonder when this is the only way of forming an extension module M. The answer will depend on the ring R and on the modules A and B. For a fixed ring, one may try to find properties of the modules A or B which ensure that *every* extension M formed by A and B is isomorphic to their direct sum. More precisely, focusing on A and B separately one may ask: what conditions on the R-module A ensure that for every R-module M,

$$\begin{gathered}\text{whenever } A \text{ is a submodule of } M\text{, there is a submodule } N \text{ of } M \\ \text{such that } M = A \oplus N,\end{gathered} \tag{$*$}$$

and what conditions on the R-module B ensure that for every R-module M,

$$\begin{gathered}\text{whenever } M/N \cong B \text{ for some submodule } N\text{, there is a submodule } B' \text{ of } M \\ \text{such that } M = N \oplus B'\end{gathered} \tag{$**$}$$

(note that in $(**)$ $B' \cong B$). Modules which have the intrinsic property $(**)$ or $(*)$ are given special names in the next definition (which phrases these properties in terms of homomorphisms).

Definition. Let R be any ring.

(1) An R-module P is called *projective* if whenever M is an R-module and $\varphi : M \to P$ is a surjective R-module homomorphism, then there is an R-submodule P' of M such that $M = \ker \varphi \oplus P'$.

(2) An R-module Q is called *injective* if whenever M is an R-module and $\psi : Q \to M$ is an injective R-module homomorphism, then there is an R-submodule N of M such that $M = Q' \oplus N$, where $Q' = \psi(Q)$.

Note that in (1), $P' \cong P$ and in (2), $Q' \cong Q$. The definition of a projective module may also be phrased: P is projective if whenever an R-module maps onto P, then the kernel is a direct summand. Similarly, the definition of an injective module may be phrased: Q is injective if it is a direct summand of every R-module that contains it as a submodule. Note that the properties "projective" and "injective" depend only on the isomorphism types of the modules P and Q, respectively.

Examples

(1) If $R = F$ is a field, every F-module is injective. This is just the Building Up Lemma (Section 11.1): if A is a subspace of the vector space M, there is a complementary subspace B of M such that $M = A \oplus B$. To see this, take a basis $\mathcal{E}_1$ of A, build it up to a basis $\mathcal{E}$ of M and let B be the subspace of M spanned by $\mathcal{E} - \mathcal{E}_1$.

By a similar argument one sees that every F-module is projective. Furthermore, irreducible F-modules are just 1-dimensional subspaces, so the fact that every F-module is completely reducible is equivalent to the statement that every vector space has a basis.

(2) Maschke's Theorem for group algebras can be stated succinctly in the following form: if the characteristic of F does not divide $|G|$, then every FG-module is injective. Note that when G has order 1, Maschke's Theorem is the statement that every F-module is injective.

(3) If R is a Principal Ideal Domain, the situation is already more complicated, even for finitely generated modules. For simplicity let $R = \mathbb{Z}$. We proved that every

finitely generated $\mathbb{Z}$-module (i.e. every finitely generated abelian group) is a direct sum of cyclic modules. However, no cyclic $\mathbb{Z}$-module A has the *intrinsic* property that whenever A appears as a submodule (subgroup) of the $\mathbb{Z}$-module M, then A is *always* a direct summand of M, i.e. $M = A \oplus N$ for some submodule N. In other words,

$$\textit{no cyclic } \mathbb{Z}\textit{-module is injective.}$$

To see this consider first the situation when $A = \mathbb{Z}$ is the infinite cyclic module. Then A appears as the proper submodule $2\mathbb{Z}$ of the module $M = \mathbb{Z}$, but M cannot be written as a direct sum of two proper submodules (i.e. M is indecomposable) — this is because every proper submodule of $\mathbb{Z}$ is of the form $n\mathbb{Z}$ for some $n \in \mathbb{Z}^+$ and any two such submodules always have nonzero intersection.

Suppose next that A is the finite cyclic module $\mathbb{Z}/n\mathbb{Z}$, $n > 1$. Then A appears as a proper submodule of the cyclic module $M \cong \mathbb{Z}/n^2\mathbb{Z}$. If A had a complement in M, it would follow that M would be isomorphic to $(\mathbb{Z}/n\mathbb{Z}) \oplus (\mathbb{Z}/n\mathbb{Z})$. The latter group, however, is not cyclic (the orders of the factors are not relatively prime), a contradiction.

As an exercise one can extend this fact about cyclic modules to direct sums to obtain more generally that

$$\textit{no finitely generated } \mathbb{Z}\textit{-module is injective.}$$

There are infinitely generated injective $\mathbb{Z}$-modules (the proof of the following is outlined in the exercises):

$$\mathbb{Q} \textit{ is an injective } \mathbb{Z}\textit{-module.}$$

The situation for *projective* $\mathbb{Z}$-modules is quite different. If M is any $\mathbb{Z}$-module which possesses a quotient module M/K isomorphic to $\mathbb{Z}$, then M always decomposes as $K \oplus P'$, where $P' \cong \mathbb{Z}$. To see this let $\varphi : M \to \mathbb{Z}$ be a surjective $\mathbb{Z}$-module homomorphism and take P' to be a cyclic submodule of M generated by any preimage of 1 (i.e. by any element of the set $\varphi^{-1}(1)$). It is easy to check that $M = \langle K, P' \rangle$ and $K \cap P' = 1$. Thus $P = \mathbb{Z}$ has the *intrinsic* property that whenever it appears as a quotient of a $\mathbb{Z}$-module M, there is a submodule of M isomorphic to P mapping onto this quotient, i.e.

$$\mathbb{Z} \textit{ is a projective } \mathbb{Z}\textit{-module.}$$

We shall classify the finitely generated projective $\mathbb{Z}$-modules shortly. In particular, we shall see that $\mathbb{Z}$ is the only cyclic $\mathbb{Z}$-module which is projective. Furthermore, a proof that $\mathbb{Q}$ is not a projective $\mathbb{Z}$-module is outlined in the exercises.

Finally, we emphasize that although every finitely generated $\mathbb{Z}$-module is a direct sum of cyclic modules, this is not a decomposition into irreducible submodules (cyclic $\mathbb{Z}$-modules are irreducible only when they are of prime order), so finitely generated $\mathbb{Z}$-modules are *not*, in general, completely reducible.

(4) We shall see in the next section that if F is any field and $n \in \mathbb{Z}^+$ then the ring $R = M_n(F)$ of all $n \times n$ matrices with entries from F has the property that every R-module is both projective and injective. As with group rings of non-abelian groups, R is not commutative and has zero divisors when $n > 1$. We shall, however, see that the ideal structure of R is especially simple.

In Example 3 above we saw that $\mathbb{Z}$ is a projective $\mathbb{Z}$-module. We now prove a generalization of this which will provide more examples of projective modules.

Proposition 4. Let R be a ring and let P be an R-module. Then the following are equivalent:

(1) P is projective

(2) whenever $\varphi : M \to P$ is a surjective homomorphism of R-modules, there is an R-module homomorphism $\psi : P \to M$ such that $\varphi \circ \psi$ is the identity on P

(3) P is a direct summand of a free R-module, i.e. there is a free R-module X such that $X = P' \oplus B$, for some R-modules B and P' of X with $P \cong P'$.

Proof: (2) $\Rightarrow$ (1): In general if $\varphi : A \to B$ and $\psi : B \to A$ are R-module homomorphisms such that $\varphi \circ \psi$ is the identity map on B, then

(i) ψ is injective

(ii) φ is surjective

(iii) $\psi(B) \cap \ker \varphi = 0$.

Furthermore, one sees that $A = \psi(B) + \ker \varphi$ by writing each element $a \in A$ as $a = \psi(\varphi(a)) + (a - \psi(\varphi(a)))$ and observing that $a - \psi(\varphi(a)) \in \ker \varphi$. Thus we also have

(iv) $A = \psi(B) \oplus \ker \varphi$.

Taking $A = M$ and $B = P$, one sees that (1) holds.

(1) $\Rightarrow$ (2): Assume P is projective and $\varphi : M \to P$ is a surjective homomorphism. By definition, there exists an R-submodule P' of M such that $M = \ker \varphi \oplus P'$. Thus $\varphi|_{P'}$ is an isomorphism from P' onto P and so $\psi = \varphi^{-1} : P \to P'$ is the desired R-module homomorphism.

(2) $\Rightarrow$ (3): Let X be any free R-module mapping onto P (for example, if S is a generating set for P let $X = F(S)$ be the free module on the set S). Let $\varphi : X \to P$ be the surjective homomorphism and let $B = \ker \varphi$. By (2) there is a homomorphism $\psi : P \to X$ such that $\varphi \circ \psi$ is the identity on P. This forces ψ to be injective and the image of ψ to intersect the kernel of φ in the zero module. Let $P' = \psi(P)$. It follows that

$$P' \cong P \qquad \text{and} \qquad X = P' \oplus B.$$

This establishes (3).

(3) $\Rightarrow$ (2): Let M be any R-module with $\varphi : M \to P$ a surjective R-module homomorphism. By assumption there is a free R-module X such that $X = P \oplus B$ for some R-module B. Form the module $M \oplus B$ and define

$$\tilde{\varphi} : M \oplus B \to P \oplus B \qquad \text{by} \qquad \tilde{\varphi}(m + b) = \varphi(m) + b.$$

Then $\tilde{\varphi}$ is a surjective R-module homomorphism from $M \oplus B$ to the free module X. Let S be a basis for X and for each $s \in S$ pick some element of $(\tilde{\varphi})^{-1}(s)$ and call it $\psi(s)$. By the universal mapping property of free modules (Theorem 6, Section 10.3) ψ extends to an R-module homomorphism from X to $M \oplus B$ (also denoted by ψ). By construction (check this on the basis S)

$$\tilde{\varphi} \circ \psi : X \to X \quad \text{is the identity map.}$$

Let $P' = \psi(P)$. By the above, ψ is injective so $P' \cong P$. It also follows from the above and the definition of $\tilde{\varphi}$ that the image of P under ψ lies in M. Thus $\psi|_P : P \to M$ is an R-module homomorphism such that $\varphi \circ \psi|_P$ is the identity map of P, i.e. (2) holds. This completes the proof.

The equivalence of (1) and (3) in Proposition 4 immediately gives:

Corollary 5. Free modules are projective.

There is an equivalence for injective modules corresponding to (1) and (2) but not to (3): the proof is left as an exercise.

Proposition 6. Let R be a ring and let Q be an R-module. Then the following are equivalent:

(1) Q is injective

(2) whenever $\varphi : Q \to M$ is an injective homomorphism of R-modules, there is an R-module homomorphism $\psi : M \to Q$ such that $\psi \circ \varphi$ is the identity on Q.

Although projective modules are not generally injective nor are injective modules projective, for certain rings these classes of modules do coincide. These rings (which by Maschke's Theorem include certain group algebras) will be the focus of our study. For such rings the following equivalence will be useful.

Proposition 7. Let R be a ring. The following are equivalent:

(1) every R-module is projective

(2) every R-module is injective.

Proof: Assume every R-module is projective and let Q be any R-module. If Q is a submodule of some module M, then by assumption $P = M/Q$ is projective. Thus there is some submodule P' of M such that $M = Q \oplus P'$. Since the submodule Q is a direct summand of every module containing it, Q is injective.

Conversely, assume every R-module is injective and let P be any R-module. If M is any R-module containing a submodule Q such that $M/Q \cong P$, then by assumption Q is injective. Thus there is some submodule P' of M such that $M = Q \oplus P'$ (and necessarily $P' \cong P$). Since for every R-module mapping onto P the kernel is a direct summand, P is projective.

EXERCISES

Let R be a ring.

1. Let P_1 and P_2 be R-modules. Prove that $P_1 \oplus P_2$ is a projective R-module if and only if both P_1 and P_2 are projective.
2. Let Q_1 and Q_2 be R-modules. Prove that $Q_1 \oplus Q_2$ is an injective R-module if and only if both Q_1 and Q_2 are injective.
3. Prove that no nonzero finitely generated $\mathbb{Z}$-module is injective.
4. Let R be a Principal Ideal Domain. Prove that a finitely generated R-module is projective if and only if it is free.

5. Let R be a Principal Ideal Domain. Prove that if a cyclic R-module is irreducible then its annihilator is a prime ideal of R.
6. Let R be a Principal Ideal Domain. Prove that if a cyclic R-module is indecomposable then its annihilator is either zero or is generated by a power of a prime in R.
7. Prove that a finitely generated $\mathbb{Z}$-module is completely reducible if and only if it is a finite abelian group in which all the Sylow subgroups are elementary abelian.

For the next two exercises recall that a nonzero $\mathbb{Z}$-module (i.e. an abelian group) Q is called *divisible* if $Q = nQ$ for all nonzero integers n. The rational numbers, $\mathbb{Q}$, are an example of a divisible $\mathbb{Z}$-module.

8. Let Q be a divisible $\mathbb{Z}$-module. Prove that Q is not a projective $\mathbb{Z}$-module. [Suppose to the contrary that Q is projective, so there is a $\mathbb{Z}$-module B such that $F = Q \oplus B$ is a free $\mathbb{Z}$-module. Use a basis of F to show that $\cap_{n=1}^{\infty} nF = 0$ and derive a contradiction from this.]
9. Let Q be a divisible $\mathbb{Z}$-module. Prove that Q is an injective $\mathbb{Z}$-module as follows: suppose Q is a submodule of the $\mathbb{Z}$-module M.
   (a) Prove that there is a $\mathbb{Z}$-submodule A of M that is maximal (under inclusion) with respect to the property $Q \cap A = 0$. [Use Zorn's Lemma.]
   (b) Show that $M = Q + A$ and deduce that $M = Q \oplus A$, where A is the submodule in (a). [Let $\overline{M} = M/A$ and assume $\overline{M} \neq \overline{Q}$. Use the maximality of A to prove that every nonzero submodule of $\overline{M}$ intersects $\overline{Q}$ nontrivially. Show then that if $x \in \overline{M} - \overline{Q}$ then $nx \in \overline{Q}$ for some positive integer n. Use the divisibility of Q to show that $\overline{M}$ contains an element y such that $\mathbb{Z}y \cap \overline{Q} = 0$, a contradiction.]
10. Let R be a commutative ring with 1 and let M be a finitely generated R-module. Recall that $\mathrm{Hom}_R(M, R)$ is also a left R-module (cf. Section 10.2). Prove that if M is a projective R-module then so is $\mathrm{Hom}_R(M, R)$.

15.3 STATEMENT OF WEDDERBURN'S THEOREM AND SOME CONSEQUENCES

In this section we state without proof a famous classification theorem due to Wedderburn which will describe, in particular, the structure of the group algebra FG when the characteristic of F does not divide the order of G. From this classification theorem we shall derive various consequences, including the fact that for each finite group G there are only a finite number of nonisomorphic irreducible FG-modules. This result, together with Maschke's Theorem, in some sense completes the Hölder Program for representation theory of finite groups over such fields. The remainder of the book is concerned with developing techniques for determining and working with the irreducible representations as well as applying this knowledge to obtain group-theoretic information.

Theorem 8. (Wedderburn) Let R be a nonzero ring with 1 (not necessarily commutative). Then the following are equivalent:
(1) every R-module is projective
(2) every R-module is injective

(3) every R-module is completely reducible

(4) the ring R considered as a left R-module is a direct sum:

$$R = L_1 \oplus L_2 \oplus \cdots \oplus L_n$$

where each L_i is a simple module (i.e. a simple left ideal) with $L_i = Re_i$, for some $e_i \in R$ with

(i) $e_i e_j = 0$ if $i \neq j$

(ii) $e_i^2 = e_i$ for all i

(iii) $\sum_{i=1}^{n} e_i = 1$

(5) as rings, R is isomorphic to a direct product of matrix rings over division rings, i.e. $R = R_1 \times R_2 \times \cdots \times R_r$ where R_j is a two-sided ideal of R and R_j is isomorphic to the ring of all $n_j \times n_j$ matrices with entries in some division ring Δ_j, $j = 1, 2, \ldots, r$.

Proof: Proofs of Wedderburn's Theorem may be found in standard references on representation theory such as *Representation Theory of Finite Groups and Associative Algebras* by C. Curtis and I. Reiner, John Wiley & Sons, 1966. The particular set of equivalences above appears in *Rings and Homology* by J. Jans, Holt, Rinehart and Winston, 1964.

Definition. A ring R satisfying any of the (equivalent) properties in Theorem 8 is called *semisimple with minimum condition.*

Rings R satisfying any of the equivalent conditions of Theorem 8 also satisfy the *minimum condition* or *descending chain condition (D.C.C)* on left ideals:

$$\text{if } I_1 \supseteq I_2 \supseteq \cdots \quad \text{is a descending chain of left ideals of } R$$
$$\text{then there is an } N \in \mathbb{Z}^+ \text{ such that } I_k = I_N \text{ for all } k \geq N$$

(which explains the use of this term in the definition above). The rings we deal with will all have this minimum condition. For example, group algebras always have this property since in any strictly descending chain of ideals the vector space dimension of the ideals (which are F-subspaces of FG) is strictly decreasing, hence the length of a strictly descending chain is at most the dimension of FG $(= |G|)$. We shall therefore use the term "semisimple" to mean "semisimple with minimum condition." The rings R_i in conclusion (5) of Wedderburn's Theorem are called the *Wedderburn components* of R and the direct product decomposition of R is called its *Wedderburn decomposition.*

The proof of Wedderburn's Theorem would take a few pages, and although it is within the scope of this book, we refer the reader to the books listed above rather than interrupt our introduction to the theory at this point. Some of the implications (e.g. (1) $\Leftrightarrow$ (2)) have already been proved; some others appear in the exercises.

One should note that condition (5) is a two-sided condition which describes the overall structure of R completely (the ring operations in this direct product of rings are componentwise addition and multiplication). In particular it implies that a semisimple ring also has the minimum condition on right ideals. A useful way of thinking of the elements of the direct product $R_1 \times \cdots \times R_r$ in conclusion (5) is as $n \times n$ (block

diagonal) matrices of the form

$$\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_r \end{pmatrix}$$

where A_i is an arbitrary $n_i \times n_i$ matrix with entries from Δ_i (here $n = \sum_{i=1}^{r} n_i$).

Recall from the preceding section that an R-module Q is *injective* if whenever Q is a submodule of any R-module M, then M has a submodule N such that $M = Q \oplus N$. Maschke's Theorem therefore implies:

Corollary 9. If G is a finite group and F is a field whose characteristic does not divide $|G|$, then the group algebra FG is a semisimple ring.

Before obtaining more precise information about how the invariants n, r, Δ_j, etc. relate to invariants in group rings FG for certain fields F, we first study the structure of matrix rings (i.e. the rings described in conclusions (4) and (5) of Wedderburn's Theorem). We introduce some terminology which is used extensively in ring theory. Recall that the *center* of the ring R is the subring of elements commuting with all elements in R; it will be denoted by $Z(R)$ (the center will contain 1 if the ring has a 1).

Definition

(1) A nonzero element e in a ring R is called an *idempotent* if $e^2 = e$.

(2) Idempotents e_1 and e_2 are said to be *orthogonal* if $e_1 e_2 = e_2 e_1 = 0$.

(3) An idempotent e is said to be *primitive* if it cannot be written as a sum of two (commuting) orthogonal idempotents.

(4) The idempotent e is called a *primitive central idempotent* if $e \in Z(R)$ and e cannot be written as a sum of two orthogonal idempotents in the ring $Z(R)$.

Proposition 10 describes the ideal structure of a matrix ring and Proposition 12 extends these results to direct products of matrix rings.

Proposition 10. Let Δ be a division ring, let $n \in \mathbb{Z}^+$, let R be the ring of all $n \times n$ matrices with entries from Δ and let I be the identity matrix (= the 1 of R).

(1) The only two-sided ideals of R are 0 and R.

(2) The center of R consists of the scalar matrices αI, where α is in the center of Δ: $Z(R) = \{\alpha I \mid \alpha \in Z(\Delta)\}$, and this is a field isomorphic to $Z(\Delta)$. In particular, if Δ is a field, the center of R is the subring of all scalar matrices. The only central idempotent in R is I (in particular, I is primitive).

(3) Let e_i be the matrix with a 1 in position i, i and zeros elsewhere. Then $e_1, \ldots, e_n$ are orthogonal primitive idempotents and $\sum_{i=1}^{n} e_i = I$.

(4) $L_i = Re_i$ is the left ideal consisting of arbitrary entries in column i and zeros in all other columns. L_i is a simple left R-module. Every simple left R-module is

isomorphic to L_1 (in particular, all L_i are isomorphic R-modules) and as a left R-module we have $R = L_1 \oplus \cdots \oplus L_n$.

Before proving this proposition it will be useful to have the following result.

Lemma 11. Let R be an arbitrary nonzero ring.

(1) If M and N are simple R-modules and $\varphi : M \to N$ is a nonzero R-module homomorphism, then φ is an isomorphism.

(2) (Schur's Lemma) If M is a simple R-module, then $\mathrm{Hom}_R(M, M)$ is a division ring.

Proof of Lemma 11: To prove (1) note that since φ is nonzero, $\ker \varphi$ is a proper submodule of M. By simplicity of M we have $\ker \varphi = 0$. Similarly, the image of φ is a nonzero submodule of the simple module N, hence $\varphi(M) = N$. This proves φ is bijective, so (1) holds.

By part (1), every nonzero element of the ring $\mathrm{Hom}_R(M, M)$ is an isomorphism, hence has an inverse. This gives (2).

Proof of Proposition 10: Let A be an arbitrary matrix in R whose i, j entry is a_{ij}. Let E_{ij} be the matrix with a 1 in position i, j and zeros elsewhere. The following straightforward computations are left as exercises:

(i) $E_{ij}A$ is the matrix whose i^{th} row equals the jth row of A and all other rows are zero.

(ii) AE_{ij} is the matrix whose j^{th} column equals the ith column of A and all other columns are zero.

(iii) $E_{pq}AE_{rs}$ is the matrix whose p, s entry is a_{qr} and all other entries are zero.

To prove (1) suppose J is any nonzero 2-sided ideal of R and let A be an element of J with a nonzero entry in position q, r. Given any $p, s \in \{1, \ldots, n\}$ we obtain from (iii) that

$$E_{ps} = \frac{1}{a_{qr}} E_{pq}AE_{rs} \in J.$$

Since the Δ-linear combinations of $\{E_{ps} \mid 1 \leq p \leq n,\ 1 \leq s \leq n\}$ give all of R, it follows that $J = R$. This proves (1).

To prove (2) assume $A \in Z(R)$. Thus for all i, j we have $E_{ij}A = AE_{ij}$. From (i) and (ii) above it follows immediately that all off-diagonal entries of A are zero and all diagonal entries of A are equal. Thus $A = \alpha I$ for some $\alpha \in \Delta$. Furthermore, A must also commute with the set of all scalar matrices βI, $\beta \in \Delta$, i.e. α must commute with all elements of Δ. Finally, since $Z(R)$ is a field, it is immediate that it contains a unique idempotent (namely I). This establishes all parts of (2).

In part (3) it is clear that $e_1, \ldots, e_n$ are orthogonal idempotents whose sum is I. We defer proving that they are primitive until we have established (4).

Next we prove (4). From (ii) above it follows that $Re_i = RE_{ii}$ is the set of matrices with arbitrary entries in the ith column and zeros in all other columns. Furthermore, if A is any nonzero element of Re_i, then certainly $RA \subseteq Re_i$. The reverse

inclusion holds because if a_{pi} is a nonzero entry of A, then by (i) above

$$e_i = E_{ii} = \frac{1}{a_{pi}} E_{ip} A \in RA.$$

This proves $Re_i = RA$ for any nonzero element $A \in Re_i$, and so Re_i must be a simple R-module.

Let M be any simple R-module. Since $Im = m$ for all $m \in M$ and since $I = \sum_{i=1}^{n} e_i$, there exists some i and some $m \in M$ such that $e_i m \neq 0$. For this i and m the map $re_i \mapsto re_i m$ is a nonzero R-module homomorphism from the simple R-module Re_i to the simple module M. By Lemma 11(1) it is an isomorphism. By (ii), the map $r \mapsto rE_{i1}$ gives $Re_i \cong Re_1$. Finally, every matrix is the direct sum of its columns so $R = L_1 \oplus \cdots \oplus L_n$. This completes the proof of (4).

It remains to prove that the idempotents in part (3) are primitive. If $e_i = a + b$, for some orthogonal idempotents a and b, then we shall see that

$$L_i = Re_i = Ra \oplus Rb.$$

This will contradict the fact that L_i is a simple R-module. To establish the above direct sum note first that since $ab = ba = 0$, we have $ae_i = a \in Re_i$ and $be_i = b \in Re_i$. For all $r \in R$ we have $re_i = ra + rb$, hence $Re_i = Ra + Rb$. Moreover, $Ra \cap Rb = 0$ because if $ra = sb$ for some $r, s \in R$, then $ra = raa = sba = 0$ (recall $a = a^2$ and $ba = 0$). This completes all parts of the proof.

Proposition 12. Let $R = R_1 \times R_2 \times \cdots \times R_r$, where R_i is the ring of $n_i \times n_i$ matrices over the division ring Δ_i, for $i = 1, 2, \ldots, r$.

(1) Identify R_i with the ith component of the direct product. Let z_i be the r-tuple with the identity of R_i in position i and zero in all other positions. Then $R_i = z_i R$ and for any $a \in R_i$, $z_i a = a$ and $z_j a = 0$ for all $j \neq i$. The elements $z_1, \ldots, z_r$ are all of the primitive central idempotents of R. They are pairwise orthogonal and $\sum_{i=1}^{r} z_i = 1$.

(2) Let N be any left R-module and let $z_i N = \{z_i x \mid x \in N\}$, $1 \leq i \leq r$. Then $z_i N$ is a left R-submodule of N, each $z_i N$ is an R_i-module on which R_j acts trivially for all $j \neq i$, and

$$N = z_1 N \oplus z_2 N \oplus \cdots \oplus z_r N.$$

(3) The simple R-modules are the simple R_i-modules on which R_j acts trivially for $j \neq i$ in the following sense. Let M_i be the unique simple R_i-module (cf. Proposition 10). We may consider M_i as an R-module by letting R_j act trivially for all $j \neq i$. Then $M_1, \ldots, M_r$ are pairwise nonisomorphic simple R-modules and any simple R-module is isomorphic to one of $M_1, \ldots, M_r$. Explicitly, the R-module M_i is isomorphic to the simple left ideal $(0, \ldots, 0, L^{(i)}, 0, \ldots, 0)$ of all elements of R whose i^{th} component, $L^{(i)}$, consists of matrices with arbitrary entries in the first column and zeros elsewhere.

(4) For any R-module N the R-submodule $z_i N$ is a direct sum of simple R-modules, each of which is isomorphic to the module M_i in (3). In particular, if M is a simple R-module, then there is a unique i such that $z_i M = M$ and for this index i we have $M \cong M_i$; for all $j \neq i$, $z_j M = 0$.

(5) If each Δ_i equals the field F, then R is a vector space over F of dimension $\sum_{i=1}^{r} n_i^2$ and $\dim_F Z(R) = r$.

Proof: In part (1) since multiplication in the direct product of rings is componentwise it is clear that z_i times the element $(a_1, \ldots, a_r)$ of R is the r-tuple with a_i in position i and zeros elsewhere. Thus $R_i = z_i R$, z_i is the identity in R_i and $z_i a = 0$ if $a \in R_j$ for any $j \neq i$. It is also clear that $z_1, \ldots, z_r$ are pairwise orthogonal central idempotents whose sum is the identity of R. The central idempotents of R are, by definition, the idempotents in $Z(R) = F_1 \times F_2 \times \cdots \times F_r$, where F_i is the center of R_i. By Proposition 10, F_i is the field $Z(\Delta_i)$. If $w = (w_1, \ldots, w_r)$ is any central idempotent then $w_i \in F_i$ for all i, and since $w^2 = w$ we have $w_i^2 = w_i$ in the field F_i. Since 0 and 1 are the only solutions to $x^2 = x$ in a field, the only central idempotents in R are r-tuples whose entries are 0's and 1's. Thus $z_1, \ldots, z_r$ are primitive central idempotents and since every central idempotent is a sum of these, they are the complete set of primitive central idempotents of R. This proves (1).

To prove (2) let N be any left R-module. First note that for any $z \in Z(R)$ the set $\{zx \mid x \in N\}$ is an R-submodule of N. In particular, $z_i N$ is an R-submodule. Let $z_i x \in z_i N$ and let $a \in R_j$ for some $j \neq i$. By (1) we have that $a = a z_j$ and so $a z_i x = (a z_j)(z_i x) = a z_i z_j x = 0$ because $z_i z_j = 0$. Thus the R-submodule $z_i N$ is acted on trivially by R_j for all $j \neq i$. For each $x \in N$ we have by (1) that $x = 1x = z_1 x + \cdots + z_r x$, hence $N = z_1 N + \cdots + z_r N$. Finally, this sum is direct because if, for instance, $x \in z_1 N \cap (z_2 N + \cdots + z_r N)$, then $x = z_1 x$ whereas z_1 times any element of $z_2 N + \cdots + z_r N$ is zero. This proves (2).

In part (3) first note that an R_i-module M becomes an R-module when R_j is defined to act trivially on M for all $j \neq i$. For such a module M the R-submodules are the same as the R_i-submodules. Thus M_i is a simple R-module for each i since it is a simple R_i-module.

Next, let M be a simple R-module. By (2), $M = z_1 M \oplus \cdots \oplus z_r M$. Since M has no nontrivial proper R-submodules, there must be a unique i such that $M = z_i M$ and $z_j M = 0$ for all $j \neq i$. Thus the simple R-module M is annihilated by R_j for all $j \neq i$. This implies that the R-submodules of M are the same as the R_i-submodules of M, so M is therefore a simple R_i-module. By Proposition 10, M is isomorphic as an R_i-module to M_i. Since R_j acts trivially on both M and M_i for all $j \neq i$, it follows that the R_i-module isomorphism may be viewed as an R-module isomorphism as well.

Suppose $i \neq j$ and suppose $\varphi : M_i \to M_j$ is an R-module isomorphism. If $s_i \in M_i$ then $s_i = z_i s_i$ so

$$\varphi(s_i) = \varphi(z_i s_i) = z_i \varphi(s_i) = 0,$$

since $\varphi(s_i) \in M_j$ and z_i acts trivially on M_j. This contradicts the fact that φ is an isomorphism and proves that $M_1, \ldots, M_r$ are pairwise nonisomorphic simple R-modules.

Finally, the left ideal of R described in (3) is acted on trivially by R_j for all $j \neq i$ and, by Proposition 10, it is up to isomorphism the unique simple R_i-module. This left ideal is therefore a simple R-module which is isomorphic to M_i. This proves (3).

For part (4) we have already proved that if M is any simple R-module then there is a unique i such that $z_i M = M$ and $z_j M = 0$ for all $j \neq i$. Furthermore, we have shown that for this index i the simple R-module M is isomorphic to M_i. Now let N

be any R-module. Then z_iN is a module over R_i which is acted on trivially by R_j for all $j \neq i$. By Wedderburn's Theorem z_iN is a direct sum of simple R-modules. Since each of these simple summands is acted on trivially by R_j for all $j \neq i$, each is isomorphic to M_i. This proves (4).

In part (5) if each Δ_i equals the field F, then as an F-vector space

$$R \cong M_{n_1}(F) \oplus M_{n_2}(F) \oplus \cdots \oplus M_{n_r}(F).$$

Each matrix ring $M_{n_i}(F)$ has dimension n_i^2 over F, hence R has dimension $\sum_{i=1}^r n_i^2$ over F. Furthermore, the center of each $M_{n_i}(F)$ is 1-dimensional (since by Proposition 10(2) it is isomorphic to F), hence $Z(R)$ has dimension r over F. This completes the proof of the proposition.

We now apply Wedderburn's Theorem (and the above ring-theoretic calculations) to the group algebra FG. First of all, in order to apply Wedderburn's Theorem we need the characteristic of F not to divide $|G|$. In fact, since we shall be dealing with numerical data in the sections to come it will be convenient to have the characteristic of F equal to 0. Secondly, it will simplify matters if we force all the division rings which will appear in the Wedderburn decompositon of FG to equal the field F — we shall prove that imposing the condition that F be algebraically closed is sufficient to ensure this. To simplify notation we shall therefore take $F = \mathbb{C}$ for most of the remainder of the text. The reader can easily check that any algebraically closed field of characteristic 0 (e.g. the field of all algebraic numbers) can be used throughout in place of $\mathbb{C}$.

By Corollary 9 the ring $\mathbb{C}\,G$ is semisimple so by Wedderburn's Theorem

$$\mathbb{C}\,G = R_1 \times R_2 \times \cdots \times R_r$$

where R_i is the ring of $n_i \times n_i$ matrices over some division ring Δ_i. Thinking of the elements of this direct product as $n \times n$ block matrices ($n = \sum_{i=1}^r n_i$) where the ith block has entries from Δ_i, the field $\mathbb{C}$ appears in this direct product as scalar matrices and is contained in the center of $\mathbb{C}\,G$. Note that each Δ_i is a vector space over $\mathbb{C}$ of dimension $\leq n$. The next result shows that this implies each $\Delta_i = \mathbb{C}$.

Proposition 13. If Δ is a division ring that is a finite dimensional vector space over an algebraically closed field F and $F \subseteq Z(\Delta)$, then $\Delta = F$.

Proof: Since $F \subseteq Z(\Delta)$, for each $\alpha \in \Delta$ the division ring generated by α and F is a field. Also, since Δ is finite dimensional over F the field $F(\alpha)$ is a finite extension of F. Because F is algebraically closed it has no nontrivial finite extensions, hence $F(\alpha) = F$ for all $\alpha \in \Delta$, i.e. $\Delta = F$.

This proposition proves that each R_i in the Wedderburn decomposition of $\mathbb{C}\,G$ is a matrix ring over $\mathbb{C}$:

$$R_i = M_{n_i}(\mathbb{C}).$$

Now Proposition 12(5) implies that

$$\sum_{i=1}^r n_i^2 = |G|.$$

The final application in this section is to prove that r (= the number of Wedderburn components in $\mathbb{C}G$) equals the number of conjugacy classes of G. To see this, first note that Proposition 12(5) asserts that $r = \dim_{\mathbb{C}} Z(\mathbb{C}G)$. We compute this dimension in another way.

Let $\mathcal{K}_1, \ldots, \mathcal{K}_s$ be the distinct conjugacy classes of G (recall that these partition G). For each conjugacy class $\mathcal{K}_i$ of G let

$$X_i = \sum_{g \in \mathcal{K}_i} g \quad \in \mathbb{C}G.$$

Note that X_i and X_j have no common terms for $i \neq j$, hence they are linearly independent elements of $\mathbb{C}G$. Furthermore, since conjugation by a group element permutes the elements of each class, $h^{-1}X_ih = X_i$, i.e. X_i commutes with all group elements. This proves that $X_i \in Z(\mathbb{C}G)$.

We show that the X_i's form a basis of $Z(\mathbb{C}G)$ — this will prove that $s = \dim_{\mathbb{C}} Z(\mathbb{C}G) = r$, as claimed. Since the X_i's are linearly independent it remains to show they span $Z(\mathbb{C}G)$. Let $X = \sum_{g \in G} \alpha_g g$ be an arbitrary element of $Z(\mathbb{C}G)$. Since $h^{-1}Xh = X$,

$$\sum_{g \in G} \alpha_g h^{-1}gh = \sum_{g \in G} \alpha_g g.$$

Since the elements of G form a basis of $\mathbb{C}G$ the coefficients of g in the above two sums are equal:

$$\alpha_{hgh^{-1}} = \alpha_g.$$

Since h was arbitrary, every element in the same conjugacy class of a fixed group element g has the same coefficient in X, hence X can be written as a linear combination of the X_i's.

We summarize these results in the following theorem.

Theorem 14. Let G be a finite group.

(1) $\mathbb{C}G = M_{n_1}(\mathbb{C}) \times M_{n_2}(\mathbb{C}) \times \cdots \times M_{n_r}(\mathbb{C})$.

(2) $\mathbb{C}G$ has exactly r distinct isomorphism types of irreducible modules and these have complex dimensions $n_1, n_2, \ldots, n_r$ (and so G has exactly r inequivalent irreducible complex representations of the corresponding degrees).

(3) $\sum_{i=1}^{r} n_i^2 = |G|$.

(4) r equals the number of conjugacy classes in G.

Corollary 15

(1) Let A be a finite abelian group. Every irreducible complex representation of A is 1-dimensional (i.e. is a homomorphism from A into $\mathbb{C}^\times$) and A has $|A|$ inequivalent irreducible complex representations. Furthermore, every finite dimensional complex matrix representation of A is equivalent to a representation into a group of diagonal matrices.

(2) The number of inequivalent (irreducible) degree 1 complex representations of any finite group G equals $|G/G'|$.

Proof: If A is abelian, $\mathbb{C}A$ is a commutative ring. Since a $k \times k$ matrix ring is not commutative whenever $k > 1$ we must have each $n_i = 1$. Thus $r = |A|$ (= the number of conjugacy classes of A). Since every $\mathbb{C}A$-module is a direct sum of irreducible submodules, there is a basis such that the matrices are diagonal with respect to this basis. This establishes the first part of the corollary.

For a general group G, every degree 1 representation, φ, is a homomorphism of G into $\mathbb{C}^\times$. Thus φ factors through G/G'. Conversely, every degree 1 representation of G/G' gives, by composition with the natural projection $G \to G/G'$, a degree 1 representation of G. The degree 1 representations of G are therefore precisely the irreducible representations of the abelian group G/G'. Part (2) is now immediate from (1).

Examples

(1) The irreducible complex representations of a finite abelian group A (i.e. the homomorphisms from A into $\mathbb{C}^\times$) can be explicitly described as follows: decompose A into a direct product of cyclic groups

$$A \cong C_1 \times \cdots \times C_n$$

where $|C_i| = |\langle x_i \rangle| = d_i$. Map each x_i to a (not necessarily primitive) $d_i{}^{\text{th}}$ root of 1 and extend this to all powers of x_i to give a homomorphism. Since there are d_i choices for the image of each x_i, the number of distinct homomorphisms of A into $\mathbb{C}^\times = GL_1(\mathbb{C})$ defined by this process equals $|A|$. By Corollary 15, these are all the irreducible representations of A. Note that it is necessary that the field contain the appropriate roots of 1 in order to realize these representations. An exercise below explores the irreducible representations of cyclic groups over $\mathbb{Q}$.

(2) Let $G = S_3$. By Theorem 14 the number of irreducible complex representations of G is three (= the number of conjugacy classes of S_3). Since the sum of the squares of the degrees is 6, the degrees must be 1, 1 and 2. The two degree 1 representations are immediately evident: the trivial representation and the representation of S_3 into $\{\pm 1\}$ given by mapping a permutation to its sign (i.e. $\sigma \mapsto +1$ if σ is an even permutation and $\sigma \mapsto -1$ if σ is an odd permutation). The degree 2 representation can be found by decomposing the permutation representation on 3 basis vectors (described in Section 1) into irreducibles as follows: let S_3 act on the basis vectors e_1, e_2, e_3 of a vector space V by permuting their indices. The vector $t = e_1 + e_2 + e_3$ is a nonzero fixed vector, so t spans a 1-dimensional G-invariant subspace (which is a copy of the trivial representation). By Maschke's Theorem there is a 2-dimensional G-invariant complement, I. Note that the permutation representation is not a sum of degree 1 representations: otherwise it could be represented by diagonal matrices and the permutations would commute in their action — this is impossible since the representation is faithful and G is non-abelian. Thus I cannot be decomposed further, so I affords *the* irreducible 2-dimensional representation. Indeed, I is the "augmentation" submodule described in Section 1

$$I = \{w \in V \mid w = \alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 e_3 \quad \text{with} \quad \alpha_1 + \alpha_2 + \alpha_3 = 0\}.$$

Clearly $e_1 - e_2$ and $e_2 - e_3$ are independent vectors in I, hence they form a basis for this 2-dimensional space. With respect to this basis of I we obtain a matrix representation of S_3 and, for example, this matrix representation on two elements of

S_3 is:

$$(1\ 2) \mapsto \begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad (1\ 2\ 3) \mapsto \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

(3) We decompose the regular representation over $\mathbb{C}$ of an arbitrary finite group. Recall that this is the representation afforded by the left $\mathbb{C}G$-module $\mathbb{C}G$ itself. By Theorem 14, $\mathbb{C}G$ is first of all a direct product of two-sided ideals:

$$\mathbb{C}G = M_{n_1}(\mathbb{C}) \times M_{n_2}(\mathbb{C}) \times \cdots \times M_{n_r}(\mathbb{C}).$$

Now by Proposition 10(4) each $M_{n_i}(\mathbb{C})$ decomposes further as a direct sum of n_i isomorphic simple left ideals. These left ideals give a complete set of isomorphism classes of irreducible $\mathbb{C}G$-modules. Thus the regular representation (over $\mathbb{C}$) of G decomposes as *the direct sum of all irreducible representations of G, each appearing with multiplicity equal to the degree of that irreducible representation*.

We record one additional property of $\mathbb{C}G$ which we shall prove in Section 16.2.

Theorem 16. The degree of each complex irreducible representation of a finite group G divides the order of G, i.e. in the notation of Theorem 14, each $n_i \mid |G|$ for $i = 1, 2, \ldots, r$.

In the next section we shall describe the primitive central idempotents of $\mathbb{C}G$ in terms of the group elements.

EXERCISES

Let G be a finite group.

1. Prove the following implications in Wedderburn's Theorem:
   (a) (2) $\Leftrightarrow$ (3)
   (b) (5) $\Rightarrow$ (4).
2. Prove that if R is a ring with 1 such that every R-module is free then R is a division ring.
3. Let F be a field, let $f(x) \in F[x]$ and let $R = F[x]/(f(x))$. Find necessary and sufficient conditions on the factorization of $f(x)$ in $F[x]$ so that R is a semisimple ring. When R is semisimple, describe its Wedderburn decomposition. [See Proposition 16 in Section 9.5.]
4. Let G be the cyclic group of order n and let $R = \mathbb{Q}G$. Describe the Wedderburn decomposition of R and find the number and the degrees of the irreducible representations of G over $\mathbb{Q}$. In particular, show that if $n = p$ is a prime then G has exactly one nontrivial irreducible representation over $\mathbb{Q}$ and this representation has degree $p - 1$. [Recall from the first example in Section 1 that $\mathbb{Q}G = \mathbb{Q}[x]/(x^n - 1)$. Use Proposition 16 in Section 9.5 and results from Section 13.6.]
5. Let p be a prime and let $F = \mathbb{F}_p$ be the field of order p. Let G be the cyclic group of order 3 and let $R = FG$. For each of $p = 2$ and $p = 7$ describe the Wedderburn decomposition of R and find the number and the degrees of the irreducible representations of G over F.
6. Prove that if P is a p-group for some prime p, then P has a faithful irreducible complex representation if and only if $Z(P)$ is cyclic. [Use Exercise 18 in Section 1, Theorem 6.1(2) and Example 3.]
7. Prove that if V is an irreducible FG-module and F is an algebraically closed field then $\mathrm{Hom}_{FG}(V, V)$ is isomorphic to F (as a ring).

8. Let F be a field, let $R = M_n(F)$ and let M be the unique irreducible R-module. Prove that $\mathrm{Hom}_R(M, M)$ is isomorphic to F (as a ring).
9. Find all 2-sided ideals of $M_n(\mathbb{Z})$.

15.4 CHARACTER THEORY AND THE ORTHOGONALITY RELATIONS

In general, for groups of large order the representations are difficult to compute and unwieldy if not impossible to write down. For example, a matrix representation of degree 100 involves matrices with 10,000 entries, and a number of 100×100 matrices may be required to describe the representation, even on a set of generators for the group. There are, however, some striking examples where large degree representations have been computed and used effectively. One instance of this is a construction of the simple group J_1 by Z. Janko in 1965 (the existence problem for simple groups was discussed at the end of Section 6.2). Janko was investigating certain properties of simple groups and he found that if any simple group possessed these properties, then it would necessarily have order 175,560 and would be generated by two elements. Furthermore, he proved that a hypothetical simple group with these properties must have a 7-dimensional representation over the field $\mathbb{F}_{11}$ with two generators mapping to the two matrices

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -3 & 2 & -1 & -1 & -3 & -1 & -3 \\ -2 & 1 & 1 & 3 & 1 & 3 & 3 \\ -1 & -1 & -3 & -1 & -3 & -3 & 2 \\ -1 & -3 & -1 & -3 & -3 & 2 & -1 \\ -3 & -1 & -3 & -3 & 2 & -1 & -1 \\ 1 & 3 & 3 & -2 & 1 & 1 & 3 \\ 3 & 3 & -2 & 1 & 1 & 3 & 1 \end{pmatrix}$$

(note that for any simple group S, every representation of S into $GL_n(F)$ which does not map all group elements to the identity matrix is a faithful representation, so S is isomorphic to its image in $GL_n(F)$). In particular, Janko's calculations showed that the simple group satisfying his properties was unique, if it existed. M. Ward was able to show that these two matrices do generate a subgroup of $GL_7(\mathbb{F}_{11})$ of order 175,560 and it follows that there does exist a simple group satisfying Janko's properties.

In a similar vein, S. Norton, R. Parker and J. Thackray constructed the simple group J_4 of order 86,775,571,046,077,562,880 using a 112-dimensional representation over $\mathbb{F}_2$. This group was shown to be generated by two elements and explicit matrices in $GL_{112}(\mathbb{F}_2)$ for these two generators were computed in the course of their analysis.

In 1981, R. Griess constructed the largest of the sporadic groups, the so-called *Monster*, of order

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

His proof involves calculations of automorphisms of an algebra over $\mathbb{C}$ of dimension 196,884 and leads to a construction of the Monster by means of a representation of this degree.

By analogy, in general it is difficult to write out the explicit permutations associated to a permutation representation $\varphi : G \to S_n$ for large degrees n. There are, however, numerical invariants such as the signs and the cycle types of the permutations $\pi(g)$ and these numerical invariants might be easier to compute than the permutations themselves (i.e. it may be possible to determine the cycle types of elements without actually having to write out the permutations themselves, as in the computation of Galois groups over $\mathbb{Q}$ in Section 14.8). These invariants alone may provide enough information in a given situation to carry out some analysis, such as proving that a given group is not simple (as illustrated in Section 6.2). Furthermore, the invariants just mentioned do not depend on the labelling of the set $\{1, 2, \ldots, n\}$ (i.e. they are independent of a "change of basis" in S_n) and they are the same for elements that are conjugate in G.

In this section we show how to attach numerical invariants to linear representations. These invariants depend only on the equivalence class (isomorphism type) of the representation. In other words, for each representation $\varphi : G \to GL_n(F)$ we shall attach an element of F to each matrix $\varphi(g)$ and we shall see that this number can, in many instances, be computed without knowing the matrix $\varphi(g)$. Moreover, we shall see that these invariants are independent of the similarity class of φ (i.e. are the same for a fixed $g \in G$ if the representation φ is replaced by an equivalent representation) and that they, in some sense, characterize the similarity classes of representations of G.

Throughout this section G is a finite group and, for the moment, F is an arbitrary field. All representations considered are assumed to be finite dimensional.

Definition

(1) A *class function* is any function from G into F which is constant on the conjugacy classes of G, i.e. $f : G \to F$ such that $f(g^{-1}xg) = f(x)$ for all $g, x \in G$.

(2) If φ is a representation of G afforded by the FG-module V, the *character* of φ is the function

$$\chi : G \to F \qquad \text{defined by} \qquad \chi(g) = \operatorname{tr} \varphi(g),$$

where $\operatorname{tr} \varphi(g)$ is the trace of the matrix of $\varphi(g)$ with respect to some basis of V (i.e. the sum of the diagonal entries of that matrix). The character is called *irreducible* or *reducible* according to whether the representation is irreducible or reducible, respectively. The *degree* of a character is the degree of any representation affording it.

In the notation of the second part of this definition we shall also refer to χ as the character afforded by the FG-module V. In general, a character is *not* a homomorphism from a group into either the additive or multiplicative group of the field.

Examples

(1) The character of the trivial representation is the function $\chi(g) = 1$ for all $g \in G$. This character is called the *principal* character of G.

(2) For degree 1 representations, the character and the representation are usually identified (by identifying a 1×1 matrix with its entry). Thus for abelian groups, irreducible complex representations and their characters are the same (cf. Corollary 15).

(3) Let $\Pi : G \to S_n$ be a permutation representation and let φ be the resulting linear representation on the basis $e_1, \ldots, e_n$ of the vector space V:

$$\varphi(g)(e_i) = e_{\Pi(g)(i)}$$

(cf. Example 4 of Section 1). With respect to this basis the matrix of $\varphi(g)$ has a 1 in the diagonal entry i, i if $\Pi(g)$ fixes i; otherwise, the matrix of $\varphi(g)$ has a zero in position i, i. Thus if π is the character of φ then

$$\pi(g) = \text{the number of fixed points of } g \text{ on } \{1, 2, \ldots, n\}.$$

In particular, if Π is the permutation representation obtained from left multiplication on the set of left cosets of some subgroup H of G then the resulting character is called the *permutation character* of G on H.

(4) The special case of Example 3 when Π is the regular permutation representation of G is worth recording: if φ is the regular representation of G (afforded by the module FG) and ρ is its character:

$$\rho(g) = \begin{cases} 0 & \text{if } g \neq 1 \\ |G| & \text{if } g = 1 \end{cases}$$

The character of the regular representation of G is called the *regular character* of G. Note that this provides specific examples where a character takes on the value 0 and is not a group homomorphism from G into either F or $F^\times$.

(5) Let $\varphi : D_{2n} \to GL_2(\mathbb{R})$ be the explicit matrix representation described in Example 6 in the second set of examples of Section 1. If χ is the character of φ then, by taking traces of the given 2×2 matrices one sees that $\chi(r) = 2\cos(2\pi/n)$ and $\chi(s) = 0$. Since φ takes the identity of D_{2n} to the 2×2 identity matrix, $\chi(1) = 2$.

(6) Let $\varphi : Q_8 \to GL_2(\mathbb{C})$ be the explicit matrix representation described in Example 7 in the second set of examples of Section 1. If χ is the character of φ then, by taking traces of the given 2×2 matrices, $\chi(i) = 0$ and $\chi(j) = 0$. Since the element $-1 \in Q_8$ maps to minus the 2×2 identity matrix, $\chi(-1) = -2$. Since φ takes the identity of Q_8 to the 2×2 identity matrix, $\chi(1) = 2$.

(7) Let $\varphi : Q_8 \to GL_4(\mathbb{R})$ be the matrix representation described in Example 8 in the second set of examples of Section 1. If χ is the character of φ then, by inspection of the matrices exhibited, $\chi(i) = \chi(j) = 0$. Since φ takes the identity of Q_8 to the 4×4 identity matrix, $\chi(1) = 4$.

For $n \times n$ matrices A and B, direct computation shows that $\operatorname{tr} AB = \operatorname{tr} BA$. If A is invertible, this implies that

$$\operatorname{tr} A^{-1}BA = \operatorname{tr} B.$$

Thus the character of a representation is independent of the choice of basis of the vector space affording it, i.e.

$$\textit{equivalent representations have the same character.} \tag{15.1}$$

Let φ be a representation of G of degree n with character χ. Since $\varphi(g^{-1}xg) = \varphi(g)^{-1}\varphi(x)\varphi(g)$ for all $g, x \in G$, taking traces of both sides shows that

$$\textit{the character of a representation is a class function.} \tag{15.2}$$

Since the trace of the $n \times n$ identity matrix is n and φ takes the identity of G to the identity linear transformation (or matrix),

$$\chi(1) \textit{ is the degree of } \varphi. \tag{15.3}$$

If V is an FG-module whose corresponding representation has character χ, then each element of the group ring FG acts as a linear transformation from V to V. Thus each $\sum_{g \in G} \alpha_g g \in FG$ has a trace when it is considered as a linear transformation from V to V. The trace of $g \in G$ acting on V is, by definition, $\chi(g)$. Since the trace of any linear combination of matrices is the linear combination of the traces, the trace of $\sum_{g \in G} \alpha_g g$ acting on V is $\sum_{g \in G} \alpha_g \chi(g)$. Note that this trace function on FG is the unique extension of the character χ of G to an F-linear transformation from FG to F. In this way we shall consider characters of G as also being defined on the group ring FG.

Notice in Example 3 above that if the field F has characteristic $p > 0$, the values of the character mod p might be zero even though the number of fixed points is nonzero. In order to circumvent such anomalies and to use the consequences of Wedderburn's Theorem obtained when F is algebraically closed we again specialize the field to be the complex numbers (or any algebraically closed field of characteristic 0). By the results of the previous section

$$\mathbb{C}G = M_{n_1}(\mathbb{C}) \times M_{n_2}(\mathbb{C}) \times \cdots \times M_{n_r}(\mathbb{C}). \tag{15.4}$$

For the remainder of this section fix the following notation:

$$\begin{aligned} &M_1, M_2, \ldots, M_r \text{ are the inequivalent irreducible } \mathbb{C}G\text{-modules,} \\ &\chi_i \text{ is the character afforded by } M_i, \quad 1 \le i \le r. \end{aligned} \tag{15.5}$$

Thus r is the number of conjugacy classes of G and we may relabel $M_1, \ldots, M_r$ if necessary so that the degree of χ_i is n_i for all i (which is also the dimension of M_i over $\mathbb{C}$).

Now every (finite dimensional) $\mathbb{C}G$-module M is isomorphic (equivalent) to a direct sum of irreducible modules:

$$M \cong a_1 M_1 \oplus a_2 M_2 \oplus \cdots \oplus a_r M_r, \tag{15.6}$$

where a_i is a nonnegative integer indicating the multiplicity of the irreducible module M_i in this direct sum decomposition, i.e.

$$a_i M_i = \overbrace{M_i \oplus \cdots \oplus M_i}^{a_i \text{ times}}.$$

Note that if the representation φ is afforded by the module M and $M = M_1 \oplus M_2$, then we may choose a basis of M consisting of a basis of M_1 together with a basis of M_2. The matrix representation with respect to this basis is of the form

$$\varphi(g) = \begin{pmatrix} \varphi_1(g) & 0 \\ 0 & \varphi_2(g) \end{pmatrix}$$

where φ_i is the representation afforded by M_i, $i = 1, 2$. One sees immediately that if ψ is the character of φ and ψ_i is the character of φ_i, then $\psi(g) = \psi_1(g) + \psi_2(g)$, i.e. $\psi = \psi_1 + \psi_2$. By induction we obtain:

the character of a representation is the sum of the characters of the constituents appearing in a direct sum decomposition. (15.7)

If ψ is the character afforded by the module M in (6) above, this gives

$$\psi = a_1\chi_1 + a_2\chi_2 + \cdots + a_r\chi_r. \tag{15.8}$$

Thus every (complex) character is a nonnegative integral sum of irreducible (complex) characters. Conversely, by taking direct sums of modules one sees that every such sum of characters is the character of some complex representation of G.

We next prove that the correspondence between characters and equivalence classes of complex representations is *bijective*. Let $z_1, z_2, \ldots, z_r$ be the primitive central idempotents of $\mathbb{C}G$ described in the preceding section. Since these are orthogonal (or equivalently, since they are the r-tuples in the decomposition of $\mathbb{C}G$ into a direct product of r subrings which have a 1 in one position and zeros elsewhere), $z_1, \ldots, z_r$ are $\mathbb{C}$-linearly independent elements of $\mathbb{C}G$. As above, each irreducible character χ_i is a function on $\mathbb{C}G$. By Proposition 12(3) we have

(a) if $j \neq i$ then $z_jM_i = 0$, i.e. z_j acts as the zero matrix on M_j, hence $\chi_j(z_i) = 0$, and

(b) z_i acts as the identity on M_i, hence $\chi_i(z_i) = n_i$.

Thus $\chi_1, \ldots, \chi_r$ are multiples of the dual basis to the independent set $z_1, \ldots, z_r$, hence are linearly independent functions. Now if the $\mathbb{C}G$-module M described in (6) above can be decomposed in a different fashion into irreducibles, say,

$$M \cong b_1M_1 \oplus b_2M_2 \oplus \cdots \oplus b_rM_r,$$

then we would obtain a relation

$$a_1\chi_1 + a_2\chi_2 + \cdots + a_r\chi_r = b_1\chi_1 + b_2\chi_2 + \cdots + b_r\chi_r.$$

By linear independence of the irreducible characters, $b_i = a_i$ for all $i \in \{1, \ldots, r\}$. Thus, in any decomposition of M into a direct sum of irreducibles, the multiplicity of the irreducible M_i is the same, $1 \leq i \leq r$. In particular,

two representations are equivalent if and only if they have the same character. (15.9)

This uniqueness can be seen in an alternate way. First, use Proposition 12(2) to decompose an arbitrary finite dimensional $\mathbb{C}G$-module M uniquely as

$$M = z_1M \oplus z_2M \oplus \cdots \oplus z_rM.$$

By part (4) of the same proposition, z_iM is a direct sum of simple modules, each of which is isomorphic to M_i. The multiplicity of M_i in a direct sum decomposition of z_iM is, by counting dimensions, equal to $\dfrac{\dim z_iM}{\dim M_i}$. This proves that the multiplic-

ity of M_i in any direct sum decomposition of M into simple submodules is uniquely determined.

Note that, as with decompositions of $F[x]$-modules into cyclic submodules, a $\mathbb{C}G$-module may have many direct sum decompositions into irreducibles — only the multiplicities are unique (see also the exercises). More precisely, comparing with the Jordan Canonical Form of a single linear transformation, the direct summand $a_iM_i = M_i \oplus \cdots \oplus M_i$ (a_i times) which equals the submodule z_iM is the analogue of the generalized eigenspace corresponding to a single eigenvalue. This submodule of M is unique (as is a generalized eigenspace) and is called the χ_i^{th} *isotypic component* of M. Within the χ_i^{th} isotypic component, the summands M_i are analogous to the 1-dimensional eigenspaces and, just as with the eigenspace of an endomorphism there is no unique basis for the eigenspace. If $G = \langle g \rangle$ is a finite cyclic group, the isotypic components of G are the same as the generalized eigenspaces of g.

Observe that the vector space of all (complex valued) class functions on G has a basis consisting of the functions which are 1 on a given class and zero on all other classes. There are r of these, where r is the number of conjugacy classes of G, so the dimension of the complex vector space of class functions is r. Since the number of (complex) irreducible characters of G equals the number of conjugacy classes and these are linearly independent class functions, we see that

the irreducible characters are a basis for the space of all complex class functions. (15.10)

The next step in the theory of characters is to put an Hermitian inner product structure on the space of class functions and prove that the irreducible characters form an orthonormal basis with respect to this inner product. For class functions θ and ψ define

$$(\theta, \psi) = \frac{1}{|G|} \sum_{g \in G} \theta(g)\overline{\psi(g)}$$

(where the bar denotes complex conjugation). One easily checks that $(\ ,\ )$ is Hermitian: for $\alpha, \beta \in \mathbb{C}$

(a) $(\alpha\theta_1 + \beta\theta_2, \psi) = \alpha(\theta_1, \psi) + \beta(\theta_2, \psi)$,
(b) $(\theta, \alpha\psi_1 + \beta\psi_2) = \bar{\alpha}(\theta, \psi_1) + \bar{\beta}(\theta, \psi_2)$, and
(c) $(\theta, \psi) = \overline{(\psi, \theta)}$.

Our principal aim is to show that the irreducible characters form an orthonormal basis for the space of complex class functions with respect to this Hermitian form (we already know that they are a basis). This fact will follow from the orthogonality of the primitive central idempotents, once we have explicitly determined these in the next proposition.

Proposition 17. Let $z_1, \ldots, z_r$ be the orthogonal primitive central idempotents in $\mathbb{C}G$ labelled in such a way that z_i acts as the identity on the irreducible $\mathbb{C}G$-module

M_i, and let χ_i be the character afforded by M_i. Then

$$z_i = \frac{\chi_i(1)}{|G|} \sum_{g \in G} \chi_i(g^{-1})g.$$

Proof: Let $z = z_i$ and write

$$z = \sum_{g \in G} \alpha_g g.$$

Recall from Example 4 in this section that if ρ is the regular character of G then

$$\rho(g) = \begin{cases} 0 & \text{if } g \neq 1 \\ |G| & \text{if } g = 1 \end{cases} \tag{15.11}$$

and recall from the last example in Section 3 that

$$\rho = \sum_{j=1}^{r} \chi_j(1)\chi_j. \tag{15.12}$$

To find the coefficient α_g, apply ρ to zg^{-1} and use linearity of ρ together with equation (11) to obtain

$$\rho(zg^{-1}) = \alpha_g|G|.$$

Computing $\rho(zg^{-1})$ using (12) then gives

$$\sum_{j=1}^{r} \chi_j(1)\chi_j(zg^{-1}) = \alpha_g|G|. \tag{15.13}$$

Let φ_j be the irreducible representation afforded by M_j, $1 \leq j \leq r$. Since we may consider φ_j as an algebra homomorphism from $\mathbb{C}G$ into $\mathrm{End}(M_j)$, we obtain $\varphi_j(zg^{-1}) = \varphi_j(z)\varphi_j(g^{-1})$. Also, we have already observed that $\varphi_j(z)$ is 0 if $j \neq i$ and $\varphi_i(z)$ is the identity endomorphism on M_i. Thus

$$\varphi_j(zg^{-1}) = \begin{cases} 0 & \text{if } j \neq i \\ \varphi_i(g^{-1}) & \text{if } j = i. \end{cases}$$

This proves $\chi_j(zg^{-1}) = \chi_i(g^{-1})\delta_{ij}$, where δ_{ij} is zero if $i \neq j$ and is 1 if $i = j$ (called the Kronecker delta). Substituting this into equation (13) gives $\alpha_g = \frac{1}{|G|}\chi_i(1)\chi_i(g^{-1})$. This is the coefficient of g in the statement of the proposition, completing the proof.

The orthonormality of the irreducible characters will follow directly from the orthogonality of the central primitive idempotents via the following calculation:

$$\begin{aligned} z_i\delta_{ij} &= z_i z_j \\ &= \frac{\chi_i(1)}{|G|}\frac{\chi_j(1)}{|G|} \sum_{g,h \in G} \chi_i(g^{-1})\chi_j(h^{-1})gh \\ &= \frac{\chi_i(1)}{|G|}\frac{\chi_j(1)}{|G|} \sum_{y \in G} \left[\sum_{x \in G} \chi_i(xy^{-1})\chi_j(x^{-1}) \right] y \end{aligned}$$

(to get the latter sum from the former substitute y for gh and x for h). Since the elements of G are a basis of $\mathbb{C}G$ we may equate coefficients with those of z_i found in Proposition 17 to get (the coefficient of g)

$$\delta_{ij}\frac{\chi_i(1)}{|G|}\chi_i(g^{-1}) = \frac{\chi_i(1)\chi_j(1)}{|G|^2}\sum_{x\in G}\chi_i(xg^{-1})\chi_j(x^{-1}).$$

Simplifying (and replacing g by g^{-1}) gives:

$$\delta_{ij}\frac{\chi_i(g)}{\chi_j(1)} = \frac{1}{|G|}\sum_{x\in G}\chi_i(xg)\chi_j(x^{-1}) \quad \text{for all } g \in G. \tag{15.14}$$

Taking $g = 1$ in (14) gives:

$$\delta_{ij} = \frac{1}{|G|}\sum_{x\in G}\chi_i(x)\chi_j(x^{-1}). \tag{15.15}$$

The sum on the right side would be precisely the inner product (χ_i, χ_j) if $\chi_j(x^{-1})$ were equal to $\overline{\chi_j(x)}$; this is the content of the next proposition.

Proposition 18. If ψ is any character of G then $\psi(x)$ is a sum of roots of 1 in $\mathbb{C}$ and $\psi(x^{-1}) = \overline{\psi(x)}$ for all $x \in G$.

Proof: Let φ be a representation whose character is ψ, fix an element $x \in G$ and let $|x| = k$. Since the minimal polynomial of $\varphi(x)$ divides $X^k - 1$ (hence has distinct roots), there is a basis of the underlying vector space such that the matrix of $\varphi(x)$ with respect to this basis is a diagonal matrix with k^{th} roots of 1 on the diagonal. Since $\psi(x)$ is the sum of the diagonal entries (and does not depend on the choice of basis), $\psi(x)$ is a sum of roots of 1. Moreover, if ϵ is a root of 1, $\epsilon^{-1} = \bar{\epsilon}$. Thus the inverse of a diagonal matrix with roots of 1 on the diagonal is the diagonal matrix with the complex conjugates of those roots of 1 on the diagonal. Since the complex conjugate of a sum is the sum of the complex conjugates, $\psi(x^{-1}) = \operatorname{tr}\varphi(x^{-1}) = \overline{\operatorname{tr}\varphi(x)} = \overline{\psi(x)}$.

Keep in mind that in the proof of Proposition 18 we first fixed a group element x and then chose a basis of the representation space so that $\varphi(x)$ was a diagonal matrix. It is always possible to diagonalize a single element but it is possible to *simultaneously* diagonalize all $\varphi(x)$'s if and only if φ is similar to a sum of degree 1 representations.

Combining the above proposition with equation (15) proves:

Theorem 19. (The First Orthogonality Relation for Group Characters) Let G be a finite group and let $\chi_1, \ldots, \chi_r$ be the irreducible characters of G over $\mathbb{C}$. Then with respect to the inner product $(\ ,\ )$ above we have

$$(\chi_i, \chi_j) = \delta_{ij}$$

and the irreducible characters are an orthonormal basis for the space of class functions. In particular, if θ is any class function then

$$\theta = \sum_{i=1}^{r}(\theta, \chi_i)\chi_i.$$

Proof: We have just established that the irreducible characters form an orthonormal basis for the space of class functions. If θ is any class function, write $\theta = \sum_{i=1}^{r} a_i\chi_i$, for some $a_i \in \mathbb{C}$. It follows from linearity of the Hermitian product that $a_i = (\theta, \chi_i)$, as stated.

We list without proof the Second Orthogonality Relation; we shall not require it for the applications in this book.

Theorem 20. (The Second Orthogonality Relation for Group Characters) Under the notation above, for any $x, y \in G$

$$\sum_{i=1}^{r} \chi_i(x)\overline{\chi_i(y)} = \begin{cases} |C_G(x)| & \text{if } x \text{ and } y \text{ are conjugate in } G \\ 0 & \text{otherwise.} \end{cases}$$

Definition. For θ any class function on G the *norm* of θ is $(\theta, \theta)^{1/2}$ and will be denoted by $||\theta||$.

When a class function is written in terms of the irreducible characters, $\theta = \sum \alpha_i\chi_i$, its norm is easily calculated as $||\theta|| = (\sum \alpha_i^2)^{1/2}$. It follows that

a character has norm 1 *if and only if it is irreducible.*

Finally, observe that computations of the inner product of characters θ and ψ may be simplified as follows. If $\mathcal{K}_1, \ldots, \mathcal{K}_r$ are the conjugacy classes of G with sizes $d_1, \ldots, d_r$ and representatives $g_1, \ldots, g_r$ respectively, then the value $\theta(g_i)\overline{\psi(g_i)}$ appears d_i times in the sum for (θ, ψ), once for each element of $\mathcal{K}_i$. Collecting these terms gives

$$(\theta, \psi) = \frac{1}{|G|}\sum_{i=1}^{r} d_i\theta(g_i)\overline{\psi(g_i)},$$

a sum only over representatives of the conjugacy classes of G. In particular, the norm of θ is given by

$$||\theta||^2 = (\theta, \theta) = \frac{1}{|G|}\sum_{i=1}^{r} d_i|\theta(g_i)|^2.$$

Examples

(1) Let $G = S_3$ and let π be the permutation character of degree 3 described in the examples at the beginning of this section. Recall that $\pi(\sigma)$ equals the number of elements in $\{1, 2, 3\}$ fixed by σ. The conjugacy classes of S_3 are represented by 1, $(1\ 2)$ and $(1\ 2\ 3)$ of sizes 1, 3 and 2 respectively, and $\pi(1) = 3$, $\pi((1\ 2)) = 1$, $\pi((1\ 2\ 3)) = 0$. Hence

$$\begin{aligned} ||\pi||^2 &= \frac{1}{6}(1\pi(1)^2 + 3\pi((1\ 2))^2 + 2\pi((1\ 2\ 3))^2) \\ &= \frac{1}{6}(9 + 3 + 0) \\ &= 2. \end{aligned}$$

This implies that π is a sum of two distinct irreducible characters, each appearing with multiplicity 1. Let χ_1 be the principal character of S_3, so that $\chi_1(\sigma) = \overline{\chi_1(\sigma)} = 1$ for all $\sigma \in S_3$. Then

$$\begin{aligned}(\pi, \chi_1) &= \frac{1}{6}(1\pi(1)\overline{\chi_1(1)} + 3\pi((1\ 2))\overline{\chi_1((1\ 2))} + 2\pi((1\ 2\ 3))\overline{\chi_1((1\ 2\ 3))}\ ) \\ &= \frac{1}{6}(3+3+0) \\ &= 1\end{aligned}$$

so the principal character appears as a constituent of π with multiplicity 1. This proves $\pi = \chi_1 + \chi_2$ for some irreducible character χ_2 of S_3 of degree 2 (and agrees with our earlier decomposition of this representation). This also shows that the value of χ_2 on $\sigma \in S_3$ is the number of fixed points of σ minus 1.

(2) Let $G = S_4$ and let π be the natural permutation character of degree 4 (so again $\pi(\sigma)$ is the number of fixed points of σ). The conjugacy classes of S_4 are represented by 1, $(1\ 2)$, $(1\ 2\ 3)$, $(1\ 2\ 3\ 4)$ and $(1\ 2)(3\ 4)$ of sizes 1, 6, 8, 6 and 3 respectively. Again we compute:

$$\begin{aligned}||\pi||^2 &= \frac{1}{24}(1\pi(1)^2 + 6\pi((1\ 2))^2 + 8\pi((1\ 2\ 3))^2 + 6\pi((1\ 2\ 3\ 4))^2 + 3\pi((1\ 2)(3\ 4))^2) \\ &= \frac{1}{24}(16+24+8+0+0) \\ &= 2\end{aligned}$$

so π has two distinct irreducible constituents. If χ_1 is the principal character of S_4, then

$$\begin{aligned}(\pi, \chi_1) &= \frac{1}{24}(1\pi(1) + 6\pi((1\ 2)) + 8\pi((1\ 2\ 3)) + 6\pi((1\ 2\ 3\ 4)) + 3\pi((1\ 2)(3\ 4))) \\ &= \frac{1}{24}(4+12+8+0+0) \\ &= 1.\end{aligned}$$

This proves that the degree 4 permutation character is the sum of the principal character and an irreducible character of degree 3.

(3) Let $G = D_8$, where

$$D_8 = \langle r, s \mid s^2 = r^4 = 1,\ rs = sr^{-1}\rangle.$$

The conjugacy classes of D_8 are represented by 1, s, r, r^2 and sr and have sizes 1, 2, 2, 1 and 2, respectively. Let φ be the degree 2 matrix representation of D_8 obtained as in Example 6 in Section 1 from embedding a square in $\mathbb{R}^2$:

$$\varphi(s) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ \varphi(r) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\ \varphi(r^2) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},\ \varphi(sr) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Let ψ be the character of this representation (where we consider the real matrices as a subset of the complex matrices). Again, since ψ is real valued one computes:

$$\begin{aligned}||\psi||^2 &= \frac{1}{8}(1\psi(1)^2 + 2\psi(s)^2 + 2\psi(r)^2 + 1\psi(r^2)^2 + 2\psi(sr)^2) \\ &= \frac{1}{8}(4+0+0+4+0) = 1.\end{aligned}$$

This proves the representation φ is irreducible (even if we allow similarity transformations by complex matrices).

We have seen that the sum of two characters is again a character. Specifically, if ψ_1 and ψ_2 are characters of representations φ_1 and φ_2, then $\psi_1 + \psi_2$ is the character of $\varphi_1 + \varphi_2$. As a final remark in our introduction to character theory we record without proof the corresponding result for products (a special case of which is in the exercises):

Proposition 21. If ψ_1 and ψ_2 are characters, then so is their product $\psi_1\psi_2$.

The next chapter will contain further explicit character computations as well as some applications of group characters to proving theorems about certain classes of groups.

Some Remarks on Fourier Analysis and Group Characters

This brief discussion is intended to indicate some connections of the results above with other areas of mathematics.

The theory of group representations described to this point is a special branch of an area of mathematics called Harmonic Analysis. Readers may already be familiar with the basic theory of Fourier series which also falls into this realm. We make some observations which show how representation theory for finite groups corresponds to "Fourier series" for some infinite groups (in particular, to Fourier series on the circle). To be mathematically precise one needs the Lebesgue integral to insure completeness of certain (Hilbert) spaces but readers may get the flavor of things by replacing "Lebesgue" by "Riemann."

Let G be the multiplicative group of points on the unit circle in $\mathbb{C}$:

$$G = \{z \in \mathbb{C} \mid |z| = 1\}.$$

We shall usually view G as the interval $[0, 2\pi]$ in $\mathbb{R}$ with the two end points identified, i.e. as the additive group $\mathbb{R}/2\pi\mathbb{Z}$ (the isomorphism is: the real number x corresponds to the complex number e^{ix}). Note that G has a translation invariant measure, namely the Lebesgue measure, and the measure of the circle is 2π. For finite groups, the counting measure is the translation invariant measure (so the measure of a subset H is the number of elements in that subset, $|H|$) and integrals on a finite group with respect to this counting measure are just finite sums.

The space

$$L^2(G) = \{f : G \to \mathbb{C} \mid f \text{ is measurable and } |f|^2 \text{ is integrable over } G\}$$

plays the role of the group algebra of the infinite group G. This space becomes a commutative ring with 1 under the convolution of functions: for $f, g \in L^2(G)$ the product $f * g : G \to \mathbb{C}$ is defined by

$$(f * g)(x) = \frac{1}{2\pi}\int_0^{2\pi} f(x-y)g(y)\,dy \qquad \text{for all } x \in G.$$

(Recall that for a finite group H, the group algebra is also formally the ring of $\mathbb{C}$-valued functions on H under a convolution multiplication and that these functions are written as formal sums – the element $\sum \alpha_g g \in \mathbb{C}\,G$ denotes the function which sends g to $\alpha_g \in \mathbb{C}$ for all $g \in G$.)

The complete set of continuous homomorphisms of G into $GL_1(\mathbb{C})$ is given by

$$e_n(x) = \mathrm{e}^{inx}, \quad x \in [0, 2\pi], \quad n \in \mathbb{Z}.$$

(Recall that for a finite abelian group, all irreducible representations are 1-dimensional and for 1-dimensional representations, characters and representations may be identified.)

The ring $L^2(G)$ admits an Hermitian inner product: for $f, g \in L^2(G)$

$$(f, g) = \frac{1}{2\pi} \int_0^{2\pi} f(t)\overline{g(t)}\, dt.$$

Under this inner product, $\{e_n \mid n \in \mathbb{Z}\}$ is an orthonormal basis (where the term "basis" is used in the analytic sense that these are independent and 0 is the only function orthogonal to all of them). Moreover,

$$L^2(G) = \widehat{\bigoplus_{n \in \mathbb{Z}}} E_n$$

where E_n is the 1-dimensional subspace spanned by e_n, the hat over the direct sum denotes taking the closure of the direct sum in the L^2-topology, and equality indicates equality in the L^2 sense. (Recall that the group algebra of a finite abelian group is the direct sum of the irreducible 1-dimensional submodules, each occuring with multiplicity one.) These facts imply the well known result from Fourier analysis that every square integrable function $f(x)$ on $[0, 2\pi]$ has a Fourier series

$$\sum_{n=-\infty}^{\infty} c_n e^{inx}$$

where the Fourier coefficients, c_n, are given by

$$c_n = (f, e_n) = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-int}\, dt.$$

This brief description indicates how the representation theory of finite groups extends to certain infinite groups and the results we have proved may already be familiar in the latter context. In fact, there is a completely analogous theory for arbitrary (not necessarily abelian) compact Lie groups — here the irreducible (complex) representations need not be 1-dimensional but they are all finite dimensional and $L^2(G)$ decomposes as a direct sum of them, each appearing with multiplicity equal to its degree. The emphasis (at least at the introductory level) in this theory is often on the importance of being able to represent functions as (Fourier) series and then using these series to solve other problems (e.g. solve differential equations). The underlying group provides the "symmetry" on which to build this "harmonic analysis," rather than being itself the principal object of study.

EXERCISES

Let G be a finite group. Unless stated otherwise all representations and characters are over $\mathbb{C}$.

1. Prove that for $n \times n$ matrices A and B with entries from any commutative ring, tr $AB =$ tr BA.
2. In each of (a) to (c) let ψ be the character afforded by the specified representation φ.
   (a) Let φ be the degree 2 representation of D_{10} described in Example 6 in the second set of examples in Section 1 (here $n = 5$) and show that $||\psi||^2 = 1$ (hence φ is irreducible).
   (b) Let φ be the degree 2 representation of Q_8 described in Example 7 in the second set of examples in Section 1 and show that $||\psi||^2 = 1$ (hence φ is irreducible).
   (c) Let φ be the degree 4 representation of Q_8 described in Example 8 in the second set of examples in Section 1 and show that $||\psi||^2 = 4$ (hence even though φ is irreducible over $\mathbb{R}$, φ decomposes over $\mathbb{C}$ as twice an irreducible representation of degree 2).
3. If χ is an irreducible character of G, prove that the χ-isotypic subspace of a $\mathbb{C}\,G$-module is unique.
4. Prove that if N is any irreducible $\mathbb{C}\,G$-module and $M = N \oplus N$, then M has infinitely many direct sum decompositions into two copies of N.
5. Prove that a class function is a character if and only if it is a positive integral linear combination of irreducible characters.
6. Let $\varphi : G \to GL(V)$ be a representation with character ψ. Let $W = \{v \in V \mid \varphi(g)(v) = v \text{ for all } g \in G\}$ be the subspace of V fixed pointwise by all elements of G. Prove that $\dim W = (\psi, \chi_1)$, where χ_1 is the principal character of G.
7. Assume V is a $\mathbb{C}\,G$-module on which G acts by permuting the basis $\mathcal{B} = \{e_1, \ldots, e_n\}$. Write $\mathcal{B}$ as a disjoint union of the orbits $\mathcal{B}_1, \ldots, \mathcal{B}_t$ of G on $\mathcal{B}$.
   (a) Prove that V decomposes as a $\mathbb{C}\,G$-module as $V_1 \oplus \cdots \oplus V_t$, where V_i is the span of $\mathcal{B}_i$.
   (b) Prove that if v_i is the sum of the vectors in $\mathcal{B}_i$ then the 1-dimensional subspace of V_i spanned by v_i is the unique $\mathbb{C}\,G$-submodule of V_i affording the trivial representation (in other words, any vector in V_i that is fixed under the action of G is a multiple of v_i). [Use the fact that G is transitive on $\mathcal{B}_i$. See also Exercise 8 in Section 1.]
   (c) Let $W = \{v \in V \mid \varphi(g)(v) = v \text{ for all } g \in G\}$ be the subspace of V fixed pointwise by all elements of G. Deduce that $\dim W = t =$ the number of orbits of G on $\mathcal{B}$.
8. Prove the following result (sometimes called Burnside's Lemma although its origin is with Frobenius): let G be a subgroup of S_n and for each $\sigma \in G$ let $\mathrm{Fix}(\sigma)$ denote the number of fixed points of σ on $\{1, \ldots, n\}$. Let t be the number of orbits of G on $\{1, \ldots, n\}$. Then

$$t|G| = \sum_{g \in G} \mathrm{Fix}(g).$$

   [Use the preceding two exercises.]
9. Let G be a nontrivial, transitive group of permutations on the finite set Ω and let ψ be the character afforded by the linear representation over $\mathbb{C}$ obtained from Ω (cf. Example 4 in Section 1) so $\psi(\sigma)$ is the number of fixed points of σ on Ω. Now let G act on the set $\Omega \times \Omega$ by $g \cdot (\omega_1, \omega_2) = (g \cdot \omega_1, g \cdot \omega_2)$ and let π be the character afforded by the linear representation obtained from this action.
   (a) Prove that $\pi = \psi^2$.
   (b) Prove that the number of orbits of G on $\Omega \times \Omega$ is given by the inner product (ψ, ψ). [By the preceding exercises, the number of orbits on $\Omega \times \Omega$ is equal to (π, χ_1), where χ_1 is the principal character.]

(c) Recall that G is said to be *doubly transitive* on Ω if it has precisely 2 orbits in its action on $\Omega \times \Omega$ (it always has at least 2 orbits since the diagonal, $\{(\omega, \omega) \mid \omega \in \Omega\}$, is one orbit). Prove that if G is doubly transitive on Ω then $\psi = \chi_1 + \chi_2$, where χ_1 is the principal character and χ_2 is a nonprincipal irreducible character of G.

(d) Let $\Omega = \{1, 2, \ldots, n\}$ and let $G = S_n$ act on Ω in the natural fashion. Show that the character of the associated linear representation decomposes as the principal character plus an irreducible character of degree $n - 1$.

10. Let ψ be the character of any 2-dimensional representation of a group G and let x be an element of order 2 in G. Prove that $\psi(x) = 2, 0$ or -2. Generalize this to n-dimensional representations.

11. Let χ be an irreducible character of G. Prove that for every element z in the center of G we have $\chi(z) = \epsilon\chi(1)$, where ϵ is some root of 1 in $\mathbb{C}$. [Use Schur's Lemma.]

12. Let ψ be the character of some representation φ of G. Prove that for $g \in G$ the following hold:

(a) if $\psi(g) = \psi(1)$ then $g \in \ker \varphi$, and

(b) if $|\psi(g)| = \psi(1)$ and φ is faithful then $g \in Z(G)$ (where $|\psi(g)|$ is the complex absolute value of $\psi(g)$). [Use the method of proof of Proposition 18.]

13. Let $\varphi : G \to GL(V)$ be a representation and let $\chi : G \to \mathbb{C}^\times$ be a degree 1 representation. Prove that $\chi\varphi : G \to GL(V)$ defined by $\chi\varphi(g) = \chi(g)\varphi(g)$ is a representation (note that multiplication of the linear transformation $\varphi(g)$ by the complex number $\chi(g)$ is well defined). Show that $\chi\varphi$ is irreducible if and only if φ is irreducible. Show that if ψ is the character afforded by φ then $\chi\psi$ is the character afforded by $\chi\varphi$. Deduce that the product of any irreducible character with a character of degree 1 is also an irreducible character.

The next few exercises study the notion of *algebraically conjugate* characters. These exercises may be considered as extensions of Proposition 18 and some consequences of these extensions. In particular we obtain a group-theoretic characterization of the conditions under which all irreducible characters of a group take values in $\mathbb{Q}$. Let F be the subfield of $\mathbb{C}$ of all elements that are algebraic over $\mathbb{Q}$ (the field of algebraic numbers). Thus F is the algebraic closure of $\mathbb{Q}$ contained in $\mathbb{C}$ and all the results established over $\mathbb{C}$ hold without change over F.

14. Note that since $F \subseteq \mathbb{C}$, every representation $\varphi : G \to GL_m(F)$ may also be considered as a complex representation. Prove that if φ is a representation over F that is irreducible over F, then φ is also irreducible when considered over the larger field $\mathbb{C}$ (note that this is not true if F is not algebraically closed – cf. Exercise 2(c) above). Show that the set of irreducible characters of G over F is the same as the set of irreducible characters over $\mathbb{C}$ (i.e. these are exactly the same set of class functions on G). Deduce that every complex representation is equivalent to a representation over F. [Since F is algebraically closed of characteristic 0, the irreducible characters over either F or $\mathbb{C}$ are characterized by the first orthogonality relation.]

Let $\varphi : G \to GL_m(F)$ be any representation with character ψ. Let $\mathbb{Q}(\varphi)$ denote the subfield of F generated by all the entries of the matrices $\varphi(g)$ for all $g \in G$.

15. Prove that $\mathbb{Q}(\varphi)$ is a finite extension of $\mathbb{Q}$.

Now let K be any Galois extension of $\mathbb{Q}$ containing $\mathbb{Q}(\varphi)$ and let $\sigma \in \mathrm{Gal}(K/\mathbb{Q})$. In fact, since every automorphism of K extends to an automorphism of F, we may assume σ is any

automorphism of F. The map $\varphi^\sigma : G \to GL_n(F)$ is defined by letting $\varphi^\sigma(g)$ be the $n \times n$ matrix whose entries are obtained by applying the field automorphism σ to the entries of the matrix $\varphi(g)$.

16. Prove that φ^σ is a representation. Prove also that the character of φ^σ is ψ^σ, where $\psi^\sigma(g) = \sigma(\psi(g))$.

17. Prove that φ is irreducible if and only if φ^σ is irreducible.

The representation φ^σ (or character ψ^σ) is called the *algebraic conjugate* of φ by σ (or of ψ, respectively); two representations φ_1 and φ_2 (or characters ψ_1 and ψ_2) are said to be *algebraically conjugate* if there is some automorphism σ of F such that $\varphi_1^\sigma = \varphi_2$ (or $\psi_1^\sigma = \psi_2$, respectively). Some care needs to be taken with this (standard) notation since the exponential notation usually denotes a right action whereas automorphisms of F act on the left on representations: $\varphi^{(\sigma\tau)} = (\varphi^\tau)^\sigma$. Let $\mathbb{Q}(\psi)$ be the subfield of F generated by the numbers $\psi(g)$ for all $g \in G$. Let $|G| = n$ and let ϵ be a primitive n^{th} root of 1 in F.

18. Prove that $\mathbb{Q}(\psi) \subseteq \mathbb{Q}(\epsilon)$. Deduce that $\mathbb{Q}(\psi)$ is a Galois extension of $\mathbb{Q}$ with abelian Galois group. [See Proposition 18.]

Recall from Section 14.5 that $\mathrm{Gal}(\mathbb{Q}(\epsilon)/\mathbb{Q}) \cong (\mathbb{Z}/n\mathbb{Z})^\times$, where the Galois automorphisms are given on the generator ϵ by $\sigma_a : \epsilon \mapsto \epsilon^a$, where a is an integer relatively prime to n.

19. Prove that if $\sigma_a \in \mathrm{Gal}(\mathbb{Q}(\epsilon)/\mathbb{Q})$ is the field automorphism defined above, then $\psi^{\sigma_a}(g) = \psi(g^a)$ for all $g \in G$. [Use the method of Proposition 18.]

20. Prove that if g is an element of G which is conjugate to g^a for all integers a relatively prime to n, then $\psi(g) \in \mathbb{Q}$, for every character ψ of G. [Use the preceding exercise and the fact that $\mathbb{Q}$ is the field fixed by all σ_a's.]

21. Prove that an element $g \in G$ is conjugate to g^a for all integers a relatively prime to $|G|$ if and only if g is conjugate to $g^{a'}$ for all integers a' relatively prime to $|g|$.

22. Show for any positive integer n that every character of the symmetric group S_n is rational valued (i.e. $\psi(g) \in \mathbb{Q}$ for all $g \in S_n$ and all characters ψ of S_n).

The next two exercises establish the converse to Exercise 20.

23. Prove that elements x and y are conjugate in a group G if and only if $\chi(x) = \chi(y)$ for all irreducible characters χ of G.

24. Let $g \in G$ and assume that every irreducible character of G is rational valued on g. Prove that g is conjugate to g^a for every integer a relatively prime to $|G|$. [If g is not conjugate to g^a for some a relatively prime to $|G|$ then by the preceding exercise there is an irreducible character χ such that $\chi(g) \neq \chi(g^a)$. Derive a contradiction from the hypothesis that $\chi(g) \in \mathbb{Q}$.]

25. Describe which irreducible characters of the cyclic group of order n are algebraically conjugate.

26. Prove that every irreducible character of both Q_8 and D_8 is rational valued. Prove that D_{10} has an irreducible character that is not rational valued.

16

Examples and Applications of Character Theory

16.1 CHARACTERS OF GROUPS OF SMALL ORDER

The *character table* of a finite group is the table of character values formatted as follows: list representatives of the r conjugacy classes along the top row and list the irreducible characters down the first column. The entry in the table in row χ_i and column g_j is $\chi_i(g_j)$. The character table of a finite group is unique up to a permutation of its rows and columns. It is customary to make the principal character the first row and the identity the first column and to list the characters in increasing order by degrees. In our examples we shall list the size of the conjugacy classes under each class so the entire table will have $r + 2$ rows and $r + 1$ columns (although strictly speaking, the character table is the $r \times r$ matrix of character values). This will enable one to easily check the "orthogonality of rows" using the first orthogonality relation: if the classes are represented by $g_1, \ldots, g_r$ of sizes $d_1, \ldots, d_r$ then

$$(\chi_i, \chi_j) = \frac{1}{|G|} \sum_{k=1}^{r} d_k \chi_i(g_k) \overline{\chi_j(g_k)}.$$

The second orthogonality relation says that the Hermitian product of any two distinct columns of a character table is zero (i.e. it gives an "orthogonality of columns").

A number of character tables are given in the *Atlas of Finite Groups* by Conway, Curtis, Norton, Parker and Wilson, Clarendon Press, 1985. These include the character table of the Monster simple group, M. The group M has 194 irreducible characters. The smallest degree of a nonprincipal irreducible character of M is 196883 and the largest degree is on the order of 2×10^{26}. Nonetheless, it is possible to compute the values of all these characters on all conjugacy classes of M.

For the first example of a character table let $G = \langle x \rangle$ be the cyclic group of order 2. Then G has 2 conjugacy classes and two irreducible characters and its character table is:

classes:	1	x
sizes:	1	1
χ_1	1	1
χ_2	1	-1

The characters and representations of this abelian group are the same, and the irreducible representations of any abelian group are described in Example 1 at the end of Section 15.3.

Similarly, if $G = \langle x \rangle$ is cyclic of order 3, the character table of G is

classes:	1	x	x^2
sizes:	1	1	1
χ_1	1	1	1
χ_2	1	ζ	ζ^2
χ_3	1	ζ^2	ζ

where ζ is a fixed primitive cube root of 1 (and $\zeta^2 = \overline{\zeta}$).

Next we construct the character table of S_3. Recall from Example 2 in Section 15.3 that S_3 has 3 irreducible characters whose values are described in that example and in Example 1 at the end of Section 15.4.

classes:	1	$(1\,2)$	$(1\,2\,3)$
sizes:	1	3	2
χ_1	1	1	1
χ_2	1	-1	1
χ_3	2	0	-1

Next we consider D_8, adopting the notation of Example 3 of Section 15.4. By Corollary 15, D_8 has four characters of degree 1. Also, in Example 3 we constructed an irreducible degree 2 representation. Since the sum of the squares of the degrees of these representations is 8, this accounts for all irreducible representations (or, since there are 5 conjugacy classes, there are 5 irreducible representations). If we let bars denote passage to the commutator quotient group (which is the Klein 4-group), then $\overline{1} = \overline{r^2}$. The degree 1 representations (= their characters) are computed by sending generators $\overline{s}$ and $\overline{r}$ to ± 1 (and the product class is mapped to the product of the values). Matrices for the degree 2 irreducible representation were computed in Example 3 of Section 15.4 and the character of this representation can be read directly from these matrices. The character table of D_8 is therefore

classes:	1	r^2	s	r	sr
sizes:	1	1	2	2	2
χ_1	1	1	1	1	1
χ_2	1	1	-1	1	-1
χ_3	1	1	1	-1	-1
χ_4	1	1	-1	-1	1
χ_5	2	-2	0	0	0

Now we compute the character table of the quaternion group of order 8. We use the usual presentation:

$$Q_8 = \langle i, j \mid i^4 = 1,\ i^2 = j^2,\ i^{-1}ji = j^{-1} \rangle$$

and let $k = ij$ and $i^2 = -1$. The conjugacy classes of Q_8 are represented by $1, -1, i, j$ and k of sizes 1, 1, 2, 2, and 2 respectively. Since the commutator quotient of Q_8 is the Klein 4-group, there are four characters of degree 1. The one remaining irreducible

character must have degree 2 in order that the sum of the squares of the degrees be 8. Let χ_5 be the degree 2 irreducible character of Q_8. One may check that the representation $\varphi : Q_8 \to GL_2(\mathbb{C})$ described explicitly in Example 7 in the second set of examples of Section 15.1 affords χ_5, but we show how the orthogonality relations give the values of χ_5 without knowing these explicit matrices. If φ is an irreducible representation of degree 2, by Schur's Lemma (cf. Exercise 18 in Section 15.1) $\varphi(-1)$ is a 2×2 scalar matrix. Since -1 has order 2 in Q_8 it follows that $\varphi(-1)$ is $\pm$ the identity matrix, hence $\chi_5(-1) = \pm 2$. Let $\chi_5(i) = a$, $\chi_5(j) = b$ and $\chi_5(k) = c$. The orthogonality relations give

$$1 = (\chi_5, \chi_5) = \frac{1}{8}(2^2 + (\pm 2)^2 + 2a\bar{a} + 2b\bar{b} + 2c\bar{c}).$$

Since $a\bar{a}$, $b\bar{b}$ and $c\bar{c}$ are nonnegative real numbers, they must all be zero. Also, since χ_5 is orthogonal to the principal character we get:

$$0 = (\chi_1, \chi_5) = \frac{1}{8}(2 + (\pm 2) + 0 + 0 + 0),$$

hence $\chi_5(-1) = -2$. The complete character table of Q_8 is

classes:	1	-1	i	j	k
sizes:	1	1	2	2	2
χ_1	1	1	1	1	1
χ_2	1	1	-1	1	-1
χ_3	1	1	1	-1	-1
χ_4	1	1	-1	-1	1
χ_5	2	-2	0	0	0

Observe that D_8 and Q_8 have the same character table, hence

nonisomorphic groups may have the same character table.

Note that the values of the degree 2 representation of Q_8 could also have been easily calculated by applying the second orthogonality relation to each column of the character table. We leave this check as an exercise. Also note that although the degree 2 irreducible characters of D_8 and Q_8 have the same (real number) values the degree 2 representation of D_8 may be realized by real matrices whereas it may be shown that Q_8 has no faithful 2-dimensional representation over $\mathbb{R}$ (cf. Exercise 10 in Section 15.1).

For the next example we construct the character table of S_4. The conjugacy classes of S_4 are represented by 1, $(1\,2)$, $(1\,2\,3)$, $(1\,2\,3\,4)$ and $(1\,2)(3\,4)$ with sizes 1, 6, 8, 6, and 3 respectively. Since $S_4' = A_4$, there are two characters of degree 1: the principal character and the character whose values are the sign of the permutation.

To obtain a degree 2 irreducible character let V be the normal subgroup of order 4 generated by $(1\,2)(3\,4)$ and $(1\,3)(2\,4)$. Any representation φ of $S_4/V \cong S_3$ gives, by composition with the natural projection $S_4 \to S_4/V$, a representation of S_4; if the

former is irreducible, so is the latter. Let φ be the composition of the projection with the irreducible 2-dimensional representation of S_3, and let χ_3 be its character. The classes of 1 and $(1\,2)(3\,4)$ map to the identity in the S_3 quotient, $(1\,2)$ and $(1\,2\,3\,4)$ map to transpositions and $(1\,2\,3)$ maps to a 3-cycle. The values of χ_3 can thus be read directly from the values of the character of degree 2 in the table for S_3.

Since S_4 has 5 irreducible characters and the sum of the squares of the degrees is 24, there must be two remaining irreducible characters, each of degree 3. In Example 2 of Section 15.4 one of these was calculated, call it χ_4. Recall that

$$\chi_4(\sigma) = (\text{the number of fixed points of } \sigma) - 1.$$

The remaining irreducible character, χ_5, is $\chi_4\chi_2$. One can either use Proposition 21 in Section 15.4 or Exercise 13 in Section 15.4 to see that this product is indeed a character. The first orthogonality relation verifies that it is irreducible. This completes the character table for S_4:

classes:	1	$(1\,2)$	$(1\,2\,3)$	$(1\,2\,3\,4)$	$(1\,2)(3\,4)$
sizes:	1	6	8	6	3
χ_1	1	1	1	1	1
χ_2	1	-1	1	-1	1
χ_3	2	0	-1	0	2
χ_4	3	1	0	-1	-1
χ_5	3	-1	0	1	-1

From the character table of S_4 one can easily compute the character table of A_4. Note that A_4 has 4 conjugacy classes. Also $|A_4 : A_4'| = 3$, so A_4 has three characters of degree 1 with $V = A_4'$ in the kernel of each degree 1 representation. The remaining irreducible character must have degree 3. One checks directly from the orthogonality relation applied in A_4 that the character χ_4 of S_4 restricted to A_4 $(= \chi_5|_{A_4})$ is irreducible. This irreducibility check is really necessary since an irreducible representation of a group need not restrict to an irreducible representation of a subgroup (for instance, the irreducible degree 2 representation of S_3 must become reducible when restricted to any proper subgroup, since these are all abelian). The character table of A_4 is therefore seen to be:

classes:	1	$(1\,2)(3\,4)$	$(1\,2\,3)$	$(1\,3\,2)$
sizes:	1	3	4	4
χ_1	1	1	1	1
χ_2	1	1	ζ	ζ^2
χ_3	1	1	ζ^2	ζ
χ_4	3	-1	0	0

where ζ is a primitive cube root of 1 in $\mathbb{C}$.

As a final example we construct the character table of S_5. For expedience we list the character table first:

classes:	1	(1 2)	(1 2 3)	(1 2 3 4)	(1 2 3 4 5)	(1 2)(3 4)	(1 2)(3 4 5)
sizes:	1	10	20	30	24	15	20
χ_1	1	1	1	1	1	1	1
χ_2	1	-1	1	-1	1	1	-1
χ_3	4	2	1	0	-1	0	-1
χ_4	4	-2	1	0	-1	0	1
χ_5	5	-1	-1	1	0	1	-1
χ_6	5	1	-1	-1	0	1	1
χ_7	6	0	0	0	0	-2	0

The conjugacy classes and their sizes were computed in Section 4.3. Since $|S_5 : S_5'| = 2$, there are two degree 1 characters: the principal character and the "sign" character.

The natural permutation of S_5 on 5 points gives rise to a permutation character of degree 5. As with S_4 and S_3 the orthogonality relations show that the square of its norm is 2 and it contains the principal character. Thus χ_3 is the permutation character minus the principal character (and, as with the smaller symmetric groups, $\chi_3(\sigma)$ is the number of fixed points of σ minus 1). As argued with S_4, it follows that $\chi_4 = \chi_3\chi_2$ is also an irreducible character.

To obtain χ_5 recall that S_5 has six Sylow 5-subgroups. Its action by conjugation on these gives a faithful permutation representation of degree 6. If ψ is the character of the associated linear representation, then since $\sigma \in S_5$ fixes a Sylow 5-subgroup if and only if it normalizes that subgroup, we have

$$\psi(\sigma) = \text{the number of Sylow 5-subgroups normalized by } \sigma.$$

The normalizer in S_5 of the Sylow 5-subgroup $\langle(1\,2\,3\,4\,5)\rangle$ is $\langle\,(1\,2\,3\,4\,5)\,,\ (2\,3\,5\,4)\,\rangle$ and all normalizers of Sylow 5-subgroups are conjugate in S_5 to this group. This normalizer contains only the identity, 5-cycles, 4-cycles and products of two disjoint transpositions. No other cycle type normalizes any Sylow 5-subgroup so on any other class, ψ is zero. To compute ψ on the remaining three nonidentity classes note (by inspection in S_6) that in any faithful action on 6 points the following hold: an element of order 5 must be a 5-cycle (hence fixes 1 point); any element of order 4 which fixes one point must be a 4-cycle (hence fixes 2 points); an element of order 2 which is the square of an element of order 4 fixes exactly 2 points also. This gives all the values of ψ. Now direct computation shows that

$$||\psi||^2 = 2 \qquad \text{and} \qquad (\chi_1, \psi) = 1.$$

Thus $\chi_5 = \psi - \chi_1$ is irreducible of degree 5. By the same theory as for χ_4 one gets that $\chi_6 = \chi_5\chi_2$ is another irreducible character.

Since there are 7 conjugacy classes, there is one remaining irreducible character and its degree is 6. Its values can be obtained immediately from the decomposition of

the regular character, ρ (cf. Example 3 in Section 15.3 and Example 4 in Section 15.4):

$$\chi_7 = \frac{\rho - (\chi_1 - \chi_2 - 4\chi_3 - 4\chi_4 - 5\chi_5 - 5\chi_6)}{6}.$$

A direct calculation by the orthogonality relations checks that χ_7 is irreducible. Note that the values of the character χ_7 were computed without explicitly exhibiting a representation with this character.

EXERCISES

1. Calculate the character tables of $Z_2 \times Z_2$, $Z_2 \times Z_3$ and $Z_2 \times Z_2 \times Z_2$. Explain why the table of $Z_2 \times Z_3$ contains primitive 6th roots of 1.
2. Compute the degrees of the irreducible characters of D_{16}.
3. Compute the degrees of the irreducible characters of A_5. Deduce that the degree 6 irreducible character of S_5 is not irreducible when restricted to A_5. [The conjugacy classes of A_5 are worked out in Section 4.3.]
4. Using the character tables in this section, for each of parts (a) to (d) use the first orthogonality relation to write the specified permutation character (cf. Example 3, Section 15.4) as a sum of irreducible characters:
   (a) the permutation character of the subgroup A_3 of S_3
   (b) the permutation character of the subgroup $\langle(1\ 2\ 3\ 4)\rangle$ of S_4
   (c) the permutation character of the subgroup V_4 of S_4
   (d) the permutation character of the subgroup $\langle(1\ 2\ 3), (1\ 2), (4\ 5)\rangle$ of S_5 (this subgroup is the normalizer of a Sylow 3-subgroup of S_5).
5. Assume that for any character ψ of a group, ψ^2 is also a character (where $\psi^2(g) = (\psi(g))^2$) — this is a special case of Proposition 21 in Section 15.4. Using the character tables in this section, for each of parts (a) to (e) write out the values of the square, χ^2, of the specified character χ and use the first orthogonality relation to write χ^2 as a sum of irreducible characters:
   (a) $\chi = \chi_3$, the degree 2 character in the table of S_3
   (b) $\chi = \chi_5$, the degree 2 character in the table of Q_8
   (c) $\chi = \chi_5$, the last character in the table of S_4
   (d) $\chi = \chi_4$, the second degree 4 character in the table of S_5
   (e) $\chi = \chi_7$, the last character in the table of S_5.
6. Calculate the character table of A_5.
7. Show that S_6 has an irreducible character of degree 5.
8. Calculate the character table of D_{10}. (This table contains nonreal entries.)
9. Calculate the character table of D_{12}.
10. Calculate the character table of $S_3 \times S_3$.
11. Calculate the character table of $Z_3 \times S_3$.
12. Calculate the character table of $Z_2 \times S_4$.
13. Calculate the character table of $S_3 \times S_4$.
14. Let n be an integer with $n \geq 3$. Show that every irreducible character of D_{2n} has degree 1 or 2 and find the number of irreducible characters of each degree. [The conjugacy classes of D_{2n} were found in Exercises 31 and 32 of Section 4.3 and its commutator subgroup was computed in Section 5.4.]
15. Prove that the character table is an invertible matrix. [Use the orthogonality relations.]

16. For each of A_5 and D_{10} describe which irreducible characters are algebraically conjugate (cf. the exercises in Section 15.4).

17. Let p be any prime and let P be a non-abelian group of order p^3 (up to isomorphism there are two choices for P; for odd p these were constructed when the groups of order p^3 were classified in Section 5.5). This exercise determines the character table of P and shows that both isomorphism types have the same character table (the argument includes the $p = 2$ case worked out in this section).
    (a) Prove that P has p^2 characters of degree 1.
    (b) Prove that P has $p - 1$ irreducible characters of degree p and that these together with the p^2 degree 1 characters are all the irreducible characters of P. [Use Theorem 14(3) and Theorem 16 in Section 15.3.]
    (c) Deduce that (regardless of the isomorphism type) the group P has $p^2 + p - 1$ conjugacy classes, p of which are of size 1 (i.e. are central classes) and $p^2 - 1$ of which each have size p. Deduce also that the classes of size p are precisely the nonidentity cosets of the center of P (i.e. if $x \in P - Z(P)$ then the conjugacy class of x is the set of p elements in the coset $xZ(P)$).
    (d) Prove that if χ is an irreducible character of degree p then the representation affording χ is faithful.
    (e) Fix a generator, z, of the center of P and let ϵ be a fixed primitive p^{th} root of 1 in $\mathbb{C}$. Prove that if χ is an irreducible character of degree p then $\chi(z) = \epsilon^i$ for some $i \in \{1, 2, \ldots, p-1\}$. Prove further that $\chi(x) = 0$ for all $x \in P - Z(P)$. (Note then that the degree p characters are all algebraically conjugate.) [Use the same reasoning as in the construction of the character table of Q_8.]
    (f) Prove that for each $i \in \{1, 2, \ldots, p-1\}$ there is a unique irreducible character χ_i of degree p such that $\chi_i(z) = \epsilon^i$. Deduce that the character table of P is uniquely determined, and describe it. [Recall from Section 6.1 that regardless of the isomorphism type, $P' = Z(P)$ and $P/P' \cong Z_p \times Z_p$. From this one can write out the degree 1 characters. Part (e) describes the degree p characters.]

16.2 THEOREMS OF BURNSIDE AND HALL

In this section we give a "theoretical" application of character theory: Burnside's $p^a q^b$ Theorem. We also prove Philip Hall's characterization of finite solvable groups, which is a group-theoretic proof relying on Burnside's Theorem as the first step in its induction.

Burnside's Theorem

The following result was proved by Burnside in 1904. Although purely group-theoretic proofs of it were discovered recently (see Theorem 2.8 in *Finite Groups III* by B. Huppert and N. Blackburn, Springer-Verlag, Berlin, 1982) the original proof by Burnside presented here is very accessible, elegant, and quite brief (given our present knowledge of representation theory).

Theorem 1. (Burnside) For p and q primes, every group of order $p^a q^b$ is solvable.

Before undertaking the proof of Burnside's Theorem itself we establish some results of a general nature. An easy consequence of these preliminary propositions is that the degrees of the irreducible characters of any finite group divide its order. The particular results that lead directly to the proof of Burnside's Theorem appear in Lemmas 7 and 8 later. It follows quite easily that a counterexample to Burnside's Theorem of minimal order is a non-abelian simple group, and it is these two character-theoretic lemmas that give the contradiction by proving the existence of a normal subgroup.

We first establish some basic results about algebraic integers (additional properties are given in the exercises). Rings of algebraic integers play a fundamental role in number theory.

Definition. An element $\alpha \in \mathbb{C}$ is called an *algebraic integer* if it is a root of a monic polynomial with coefficients from $\mathbb{Z}$.

Algebraic integers are generalizations of the usual integers $\mathbb{Z}$, which are roots of monic polynomials of degree 1 with coefficients from $\mathbb{Z}$.

Proposition 2. Let $\alpha \in \mathbb{C}$. The following are equivalent:

(1) α is an algebraic integer,

(2) α is algebraic over $\mathbb{Q}$ and the minimal polynomial of α over $\mathbb{Q}$ has integer coefficients, and

(3) $\mathbb{Z}[\alpha]$ is a finitely generated $\mathbb{Z}$-module (where $\mathbb{Z}[\alpha]$ is the subring of $\mathbb{C}$ generated by $\mathbb{Z}$ and α, i.e. is the ring of all $\mathbb{Z}$-linear combinations of nonnegative powers of α).

Proof: Parts (1) and (2) are equivalent by Gauss' Lemma (cf. Section 9.3).

To show (1) implies (3) assume α is a root of the monic polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in \mathbb{Z}[x]$. Thus

$$\alpha^n = -(a_{n-1}\alpha^{n-1} + \cdots + a_0)$$

and so α^n (and subsequently all higher powers of α) can be expressed as integral linear combinations of $\alpha^{n-1}, \ldots, \alpha, 1$. These elements are a finite generating set for $\mathbb{Z}[\alpha]$, so (3) holds.

To show (3) implies (1) assume $\alpha_1, \alpha_2, \ldots, \alpha_n$ are a finite generating set for $\mathbb{Z}[\alpha]$. Then for $1 = 1, 2, \ldots, n$ the elements $\alpha\alpha_i$ can be written as $\mathbb{Z}$-linear combinations of $\alpha_1, \ldots, \alpha_n$:

$$\alpha\alpha_i = \sum_{j=1}^{n} a_{ij}\alpha_j.$$

After bringing everything over to the right side these equations become

$$0 = \sum_{j=1}^{n} (a_{ij} - \delta_{ij}\alpha)\alpha_j \qquad i = 1, \ldots, n$$

where δ_{ij} is the Kronecker delta. Since these homogeneous equations have a nontrivial solution $\alpha_1, \ldots, \alpha_n$, the determinant of the coefficient matrix must be zero, so α is a

root of the polynomial $\det((a_{ij}) - xI)$. Since this polynomial is $\pm$ a monic polynomial, after a possible sign change we have (1). This completes the proof.

Corollary 3. The algebraic integers in $\mathbb{C}$ form a ring and the algebraic integers in $\mathbb{Q}$ are the elements of $\mathbb{Z}$.

Proof: Let α and β be algebraic integers. By Proposition 2(3)

$$\begin{aligned}\mathbb{Z}[\alpha] &= \mathbb{Z}\alpha_1 + \cdots + \mathbb{Z}\alpha_n \\ \mathbb{Z}[\beta] &= \mathbb{Z}\beta_1 + \cdots + \mathbb{Z}\beta_m.\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{Z}[\alpha, \beta] &= \mathbb{Z}[\alpha][\beta] \\ &= \mathbb{Z}[\alpha]\beta_1 + \cdots + \mathbb{Z}[\alpha]\beta_m \\ &= \mathbb{Z}\alpha_1\beta_1 + \mathbb{Z}\alpha_2\beta_1 + \cdots + \mathbb{Z}\alpha_n\beta_1 + \cdots\cdots + \mathbb{Z}\alpha_1\beta_m + \cdots + \mathbb{Z}\alpha_n\beta_m.\end{aligned}$$

Thus $\mathbb{Z}[\alpha, \beta]$ is a finitely generated $\mathbb{Z}$-module. Since $\mathbb{Z}$ is a Principal Ideal Domain, submodules of a finitely generated module are finitely generated (cf. Chapter 12). Thus $\mathbb{Z}[\alpha \pm \beta]$ and $\mathbb{Z}[\alpha\beta]$ are finitely generated $\mathbb{Z}$-modules since both are submodules of $\mathbb{Z}[\alpha, \beta]$. By Proposition 2(3), $\alpha \pm \beta$ and $\alpha\beta$ are algebraic integers. This proves that the algebraic integers form a ring.

It is immediate from Proposition 2(2) that the algebraic integers in $\mathbb{Q}$ are the ordinary integers. This completes the proof.

Corollary 4. For every character ψ of the finite group G, $\psi(x)$ is an algebraic integer for all $x \in G$.

Proof: By Proposition 18 in Section 15.4, $\psi(x)$ is a sum of roots of 1. Clearly each root of 1 is an algebraic integer, so the result follows immediately from Corollary 3.

We shall also need some preliminary character-theoretic lemmas before beginning the main proof. Adopt the following notation for the arbitrary finite group G: $\chi_1, \ldots, \chi_r$ are the distinct irreducible (complex) characters of G, $\mathcal{K}_1, \ldots, \mathcal{K}_r$ are the conjugacy classes of G and φ_i is an irreducible matrix representation whose character is χ_i for each i.

Proposition 5. Define the complex valued function ω_i on $\{\mathcal{K}_1, \ldots, \mathcal{K}_r\}$ for each i by

$$\omega_i(\mathcal{K}_j) = \frac{|\mathcal{K}_j|\chi_i(g)}{\chi_i(1)}$$

where g is any element of $\mathcal{K}_j$. Then $\omega_i(\mathcal{K}_j)$ is an algebraic integer for all i and j.

Proof: We first prove that if I is the identity matrix, then

$$\sum_{g \in \mathcal{K}_j} \varphi_i(g) = \omega_i(\mathcal{K}_j)I. \tag{16.1}$$

To see this let X be the left hand side of (1). As we saw in Section 15.3, each $x \in G$ acting by conjugation permutes the elements of $\mathcal{K}_j$ and so X commutes with $\varphi_i(g)$ for all g. By Schur's Lemma (Exercise 18 in Section 15.1) X is a scalar matrix:

$$X = \alpha I \qquad \text{for some } \alpha \in \mathbb{C}.$$

It remains to show that $\alpha = \omega_i(\mathcal{K}_j)$. But

$$\operatorname{tr} X = \sum_{g \in \mathcal{K}_j} \operatorname{tr} \varphi_i(g) = \sum_{g \in \mathcal{K}_j} \chi_i(g) = |\mathcal{K}_j| \chi_i(g).$$

Thus $\alpha \chi_i(1) = \operatorname{tr} X = |\mathcal{K}_j| \chi_i(g)$, as needed to establish (1).

Now let g be a fixed element of $\mathcal{K}_s$ and define a_{ijs} to be the number of ordered pairs g_i, g_j with $g_i \in \mathcal{K}_i$, $g_j \in \mathcal{K}_j$ and $g_i g_j = g$. Notice that a_{ijs} is an integer. It is independent of the choice of g in $\mathcal{K}_s$ because if $x^{-1}gx$ is a conjugate of g, every ordered pair g_i, g_j whose product is g gives rise to an ordered pair $x^{-1}g_i x$, $x^{-1}g_j x$ whose product is $x^{-1}gx$ (and vice versa).

Next we prove that for all $i, j, t \in \{1, \ldots, r\}$

$$\omega_t(\mathcal{K}_i)\omega_t(\mathcal{K}_j) = \sum_{s=1}^{r} a_{ijs}\omega_t(\mathcal{K}_s). \tag{16.2}$$

To see this note that by (1), the left hand side of (2) is the diagonal entry of the scalar matrix on the left of the following equation:

$$\begin{aligned}
\Big(\sum_{g \in \mathcal{K}_i} \varphi_t(g)\Big)\Big(\sum_{g \in \mathcal{K}_j} \varphi_t(g)\Big) &= \sum_{g_i \in \mathcal{K}_i} \sum_{g_j \in \mathcal{K}_j} \varphi_t(g_i g_j) \\
&= \sum_{s=1}^{r} \sum_{g \in \mathcal{K}_s} a_{ijs}\varphi_t(g) \\
&= \sum_{s=1}^{r} a_{ijs} \sum_{g \in \mathcal{K}_s} \varphi_t(g) && \text{(since } a_{ijs} \text{ is independent of } g \in \mathcal{K}_s) \\
&= \sum_{s=1}^{r} a_{ijs}\omega_t(\mathcal{K}_s) I && (\text{ by (1) }).
\end{aligned}$$

Comparing entries of these scalar matrices gives (2).

Now (2) implies that the subring of $\mathbb{C}$ generated by $\mathbb{Z}$ and $\omega_t(\mathcal{K}_1), \ldots, \omega_t(\mathcal{K}_r)$ is a finitely generated $\mathbb{Z}$-module for each $t \in \{1, \ldots, r\}$ (it is generated as a $\mathbb{Z}$-module by $1, \omega_t(\mathcal{K}_1), \ldots, \omega_t(\mathcal{K}_r)$). Since $\mathbb{Z}$ is a Principal Ideal Domain the submodule $\mathbb{Z}[\omega_t(\mathcal{K}_i)]$ is also a finitely generated $\mathbb{Z}$-module, hence $\omega_t(\mathcal{K}_i)$ is an algebraic integer by Proposition 2. This completes the proof.

Corollary 6. The degree of each complex irreducible representation of a finite group G divides the order of G, i.e. $\chi_i(1) \mid |G|$ for $i = 1, 2, \ldots, r$.

Proof: Under the notation of Proposition 5 and with $g_j \in \mathcal{K}_j$ we have

$$\begin{aligned}\frac{|G|}{\chi_i(1)} &= \frac{|G|}{\chi_i(1)}(\chi_i, \chi_i) \\ &= \sum_{j=1}^{r} \frac{|\mathcal{K}_j|\chi_i(g_j)\overline{\chi_i(g_j)}}{\chi_i(1)} \\ &= \sum_{j=1}^{r} \omega_i(\mathcal{K}_j)\overline{\chi_i(g_j)}.\end{aligned}$$

The right hand side is an algebraic integer and the left hand side is rational, hence is an integer. This proves the corollary.

The next two lemmas lead directly to Burnside's Theorem.

Lemma 7. If G is any group that has a conjugacy class $\mathcal{K}$ and an irreducible matrix representation φ with character χ such that $(|\mathcal{K}|, \chi(1)) = 1$, then for $g \in \mathcal{K}$ either $\chi(g) = 0$ or $\varphi(g)$ is a scalar matrix.

Proof: By hypothesis there exist $s, t \in \mathbb{Z}$ such that $s|\mathcal{K}| + t\chi(1) = 1$. Thus

$$s|\mathcal{K}|\chi(g) + t\chi(1)\chi(g) = \chi(g).$$

Divide both sides of this by $\chi(1)$ and note that by Corollary 4 and Proposition 5 both $\chi(g)$ and $\dfrac{|\mathcal{K}|\chi(g)}{\chi(1)}$ are algebraic integers, hence so is $\dfrac{\chi(g)}{\chi(1)}$. Let $a_1 = \dfrac{\chi(g)}{\chi(1)}$ and let $a_1, a_2, \ldots, a_n$ be all its algebraic conjugates over $\mathbb{Q}$ (i.e. the roots of the minimal polynomial of a_1 over $\mathbb{Q}$). Since a_1 is a sum of $\chi(1)$ roots of 1 divided by the integer $\chi(1)$, each a_i is also a sum of $\chi(1)$ roots of 1 divided by $\chi(1)$. Thus a_i has complex absolute value ≤ 1 for all i. Now $b = \prod_{i=1}^{n} a_i \in \mathbb{Q}$ and b is an algebraic integer ($\pm b$ is the constant term of the irreducible polynomial of a_1), hence $b \in \mathbb{Z}$. But

$$|b| = \prod_{i=1}^{n} |a_i| \leq 1,$$

so $b = 0, \pm 1$. Since all a_i's are conjugate, $b = 0 \Leftrightarrow a_1 = 0 \Leftrightarrow \chi(g) = 0$. Also, $b = \pm 1 \Leftrightarrow |a_i| = 1$ for all i. Thus either $\chi(g) = 0$ or $|\chi(g)| = \chi(1)$. In the former situation the lemma is established, so assume $|\chi(g)| = \chi(1)$.

Let φ_1 be a matrix representation equivalent to φ in which $\varphi_1(g)$ is a diagonal matrix:

$$\varphi_1(g) = \begin{pmatrix} \epsilon_1 & & & \\ & \epsilon_2 & & \\ & & \ddots & \\ & & & \epsilon_n \end{pmatrix}$$

Thus $\chi(g) = \epsilon_1 + \cdots + \epsilon_n$. By the triangle inequality if $\epsilon_i \neq \epsilon_j$ for any i, j, then $|\epsilon_1 + \cdots + \epsilon_n| < n = \chi(1)$. Since this is not the case we must have $\varphi_1(g) = \epsilon I$ (where

$\epsilon = \epsilon_i$ for all i). Since scalar matrices are similar only to themselves, $\varphi(g) = \epsilon I$ as well. This completes the proof.

Lemma 8. If $|\mathcal{K}|$ is a power of a prime for some nonidentity conjugacy class $\mathcal{K}$ of G, then G is not a non-abelian simple group.

Proof: Suppose to the contrary that G is a non-abelian simple group and let $|\mathcal{K}| = p^c$. Let $g \in \mathcal{K}$. If $c = 0$ then $g \in Z(G)$, contrary to a non-abelian simple group having a trivial center. As above, let $\chi_1, \ldots, \chi_r$ be all the irreducible characters of G with χ_1 the principal character and let ρ be the regular character of G. By decomposing ρ into irreducibles we obtain

$$0 = \rho(g) = 1 + \sum_{i=2}^{r} \chi_i(1)\chi_i(g). \tag{16.3}$$

If $p \mid \chi_j(1)$ for every $j > 1$ with $\chi_j(g) \neq 0$, then write $\chi_j(1) = pd_j$. In this case (3) becomes

$$0 = 1 + p\sum_j d_j\chi_j(g).$$

Thus $\sum_j d_j\chi_j(g) = -1/p$ is an algebraic integer, a contradiction. This proves there is some j such that p does not divide $\chi_j(1)$ and $\chi_j(g) \neq 0$. If φ is a representation whose character is χ_j, then φ is faithful (because G is assumed to be simple) and, by Lemma 7, $\varphi(g)$ is a scalar matrix. Since $\varphi(g)$ commutes with all matrices, $\varphi(g) \in Z(\varphi(G))$. This forces $g \in Z(G)$, contrary to G being a non-abelian simple group. The proof of the lemma is complete.

We now prove Burnside's Theorem. Let G be a group of order $p^a q^b$ for some primes p and q. If $p = q$ or if either exponent is 0 then G is nilpotent hence solvable. Thus we may assume this is not the case. Proceeding by induction let G be a counterexample of minimal order. If G has a proper, nontrivial normal subgroup N, then by induction both N and G/N are solvable, hence so is G (cf. Section 3.4 or Proposition 6.10). Thus we may assume G is a non-abelian simple group. Let $P \in Syl_p(G)$. By Theorem 8 of Chapter 4 there exists $g \in Z(P)$ with $g \neq 1$. Since $P \leq C_G(g)$, the order of the conjugacy class of g (which equals $|G : C_G(g)|$) is prime to p, i.e. is a power of q. This violates Lemma 8 and so completes the proof of Burnside's Theorem.

Philip Hall's Theorem

Recall that a subgroup of a finite group is called a *Hall subgroup* if its order and index are relatively prime. For any subgroup H of a group G a subgroup K such that $G = HK$ and $H \cap K = 1$ is called a *complement* to H in G.

Theorem 9. (P. Hall) Let G be a group of order $p_1^{\alpha_1}p_2^{\alpha_2}\cdots p_t^{\alpha_t}$ where $p_1, \ldots, p_t$ are distinct primes. If for each $i \in \{1, \ldots, t\}$ there exists a subgroup H_i of G with $|G : H_i| = p_i^{\alpha_i}$, then G is solvable.

Hall's Theorem can also be phrased: if for each $i \in \{1, \ldots, t\}$ a Sylow p_i-subgroup of G has a complement, then G is solvable. The converse to Hall's Theorem is also true — this was Exercise 33 in Section 6.1.

We shall first need some elementary lemmas.

Lemma 10. If G is solvable of order > 1, then there exists $P \trianglelefteq G$ with P a nontrivial p-group for some prime p.

Proof: This is a special case of the exercise on minimal normal subgroups of solvable groups at the end of Section 6.1. One can see this easily by letting P be a nontrivial Sylow subgroup of the last nontrivial term, $G^{(n-1)}$, in the derived series of G (where G has solvable length n). In this case $G^{(n-1)}$ is abelian so P is a characteristic subgroup of $G^{(n-1)}$, hence is normal in G.

Lemma 11. Let G be a group of order $p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}$ where $p_1, \ldots, p_t$ are distinct primes. Suppose there are subgroups H and K of G such that for each $i \in \{1, \ldots, t\}$, either $p_i^{\alpha_i}$ divides $|H|$ or $p_i^{\alpha_i}$ divides $|K|$. Then $G = HK$ and $|H \cap K| = (|H|, |K|)$.

Proof: Fix some $i \in \{1, \ldots, t\}$ and suppose first that $p_i^{\alpha_i}$ divides the order of H. Since HK is a disjoint union of right cosets of H and each of these right cosets has order equal to $|H|$, it follows that $p_i^{\alpha_i}$ divides $|HK|$. Similarly, if $p_i^{\alpha_i}$ divides $|K|$, since HK is a disjoint union of left cosets of K, again $p_i^{\alpha_i}$ divides $|HK|$. Thus $|G| \mid |HK|$ and so $G = HK$. Since

$$|HK| = \frac{|H||K|}{|H \cap K|},$$

it follows that $|H \cap K| = (|H|, |K|)$.

We now begin the proof of Hall's Theorem, proceeding by induction on $|G|$. Note that if $t = 1$ the hypotheses are trivially satisfied for any group ($H_1 = 1$) and if $t = 2$ the hypotheses are again satisfied for any group by Sylow's Theorem (H_1 is a Sylow p_2-subgroup of G and H_2 is a Sylow p_1-subgroup of G). If $t = 1$, G is nilpotent, hence solvable and if $t = 2$, G is solvable by Burnside's Theorem. Assume therefore that $t \geq 3$.

Fix i and note that by the preceding lemma, for all $j \in \{1, \ldots, t\} - \{i\}$,

$$|H_i : H_i \cap H_j| = p_j^{\alpha_j}.$$

Thus every Sylow p_j-subgroup of H_i has a complement in H_i: $H_j \cap H_i$. By induction H_i is solvable.

By Lemma 10 we may choose $P \trianglelefteq H_1$ with $|P| = p_i^a > 1$ for some $i > 1$. Since $t \geq 3$ there exists an index $j \in \{1, \ldots, t\} - \{1, i\}$. By Lemma 11

$$|H_1 \cap H_j| = p_2^{\alpha_2} \cdots p_{j-1}^{\alpha_{j-1}} p_{j+1}^{\alpha_{j+1}} \cdots p_t^{\alpha_t}.$$

Thus $H_1 \cap H_j$ contains a Sylow p_i-subgroup of H_1. Since P is a normal p_i-subgroup of H_1, P is contained in every Sylow p_i-subgroup of H_1 and so $P \leq H_1 \cap H_j$. By Lemma 11, $G = H_j H_1$ so each $g \in G$ may be written $g = h_j h_1$ for some $h_j \in H_j$ and $h_1 \in H_1$. Then

$$H_j^g = H_j^{h_j h_1} = H_j^{h_1}$$

and so

$$\bigcap_{g\in G} H_j^g = \bigcap_{h_1\in H_1} H_j^{h_1}.$$

Now $P \leq H_j$ and $P^{h_1} = P$ for all $h_1 \in H_1$. Thus

$$1 \neq P \leq \bigcap_{h_1\in H_1} H_j^{h_1}.$$

Thus $N = \cap_{g\in G} H_j^g$ is a nontrivial, proper normal subgroup of G. It follows that both N and G/N satisfy the hypotheses of the theorem (cf. the exercises in Section 3.3). Both N and G/N are solvable by induction, so G is solvable. This completes the proof of Hall's Theorem.

EXERCISES

1. Let α be a complex number that is algebraic over $\mathbb{Q}$. Prove that there is a positive integer k such that $k\alpha$ is an algebraic integer.
2. Let K be a subfield of $\mathbb{C}$ that is an extension of $\mathbb{Q}$ of finite degree n. Let $\mathcal{O}_K$ be the intersection of the ring of algebraic integers with the field K — called the *ring of integers of* K.
   (a) Prove that $\mathcal{O}_K$ contains a $\mathbb{Q}$ basis for K.
   (b) Prove that K is the quotient field of $\mathcal{O}_K$.
   (c) Prove that $\mathcal{O}_K$ is a free $\mathbb{Z}$-module of rank n.
3. Let K be a subfield of $\mathbb{C}$ and let $\mathcal{O}_K$ be the ring of integers of K. Suppose R is a subring of $\mathcal{O}_K$ with $\mathbb{Z} \subseteq R$ and with the quotient field of R equal to K. Prove that if R is a Unique Factorization Domain, then $R = \mathcal{O}_K$. [Use the method of proof of Proposition 11 in Section 9.4.]
4. Prove that the ring of integers of $\mathbb{Q}(i)$ is $\mathbb{Z}[i]$.
5. Show that every character of the symmetric group S_n is integer valued, for all n (i.e. $\psi(g) \in \mathbb{Z}$ for all $g \in S_n$ and all characters ψ of S_n). [See Exercise 22 in Section 15.4.]
6. Let G be a finite group with the property that every maximal subgroup has either prime or prime squared index. Prove that G is solvable. (The simple group $GL_3(\mathbb{F}_2)$ has the property that every maximal subgroup has index either 7 or 8, i.e. either prime or prime cubed index – cf. Section 6.2.). [Show that a minimal counterexample G is simple. Let p be the largest prime dividing $|G|$ and let P be a Sylow p-subgroup of G. If $P \trianglelefteq G$, apply induction to G/P. Otherwise let M be a maximal subgroup containing $N_G(P)$. Use Exercise 51 in Section 4.5 to show that $p = 3$ and deduce that $|G| = 2^a 3^b$.]
7. Assume G is a finite group which possesses an abelian subgroup H whose index is a power of a prime. Prove that G is solvable.
8. Repeat the preceding exercise with the word "abelian" replaced by "nilpotent."

16.3 AN INTRODUCTION TO THE THEORY OF INDUCED CHARACTERS

Let G be a finite group, let H be a subgroup of G and let φ be a representation of the subgroup H over an arbitrary field F. In this section we show how to obtain a representation of G, called the induced representation, from the representation φ of

its subgroup. We also determine a formula for the character of this induced representation, the induced character, in terms of the character of φ and we illustrate this formula by computing some induced characters in specific groups. Finally, we apply the theory of induced characters to prove that there are no simple groups of order $3^3 \cdot 7 \cdot 13 \cdot 409$, a group order which was discussed at the end of Section 6.2 in the context of the existence problem for simple groups.

The theory of induced representations and induced characters marks the beginning of more advanced representation theory. This section is intended as an introductory survey rather than as a comprehensive treatment and the results we have included were chosen to serve this purpose. Also, in many books on representation theory of finite groups, induced representations are defined in terms of tensor products. These give a natural way of forming an FG-module from an FH-module. Our development, which circumvents the use of tensor products, requires slightly more work to construct an FG-module from a given FH-module; however, it has the advantage that it generalizes directly to infinite groups where continuity considerations may play a role. The FG-modules constructed from the same FH-module by tensor products or by our methods are isomorphic (as FG-modules). Since we shall be primarily interested in the character theory of induced representations this difference will play a minimal role. Those wishing to continue the study of representation theory should become conversant with both approaches.

First we observe that it may not be possible to *extend* a representation φ of the subgroup H to a representation Φ of G in such a way that $\Phi|_H = \varphi$. For example, $A_3 \leq S_3$ and A_3 has a faithful representation of degree 1 (cf. Section 1). Since every degree 1 representation of S_3 contains $A_3 = S_3'$ in its kernel, this representation of A_3 cannot be extended to a representation of S_3. For another example of a representation of a subgroup H which cannot be extended to the whole group G take G to be any simple group and let φ be any representation of H with the property that $\ker \varphi$ is a proper, nontrivial normal subgroup of H. If φ extended to a representation Φ of G then the kernel of Φ would be a proper, nontrivial subgroup of G, contrary to G being a simple group. We shall see that the method of induced characters produces a representation Φ of G from a given representation φ of its subgroup H but that $\Phi|_H \neq \varphi$ in general (indeed, unless $H = G$ the degree of Φ will be greater than the degree of φ).

Let V be an FH-module which affords the representation φ of H. Let W be the collection of all set functions f from G to V satisfying

$$f : G \to V \qquad \text{and} \qquad f(hx) = hf(x) \text{ for all } h \in H,\ x \in G,$$

where $hf(x)$ denotes the action of h on the FH-module element $f(x)$. The set W becomes a vector space over F under pointwise addition and scalar multiplication:

$$\begin{aligned} (f_1 + f_2)(x) &= f_1(x) + f_2(x) && \text{for all } f_1, f_2 \in W,\ x \in G \\ (\alpha f)(x) &= \alpha f(x) && \text{for all } \alpha \in F,\ f \in W,\ x \in G. \end{aligned}$$

Define an action of G on W as follows:

$$gf : G \to V \qquad \text{by} \qquad (gf)(x) = f(xg) \quad \text{for all } g, x \in G,\ f \in W.$$

First note that if $f \in W$ and $g \in G$ then $gf \in W$ because $(gf)(hx) = f(hxg) = hf(xg) = h(gf)(x)$. Next we show that this definition satisfies the axioms for an action of G on W. Let $g_1, g_2 \in G$, let $f \in W$ and let $g_2 f = f_2 \in W$. Then

$$\begin{aligned}[g_1(g_2 f)](x) = (g_1 f_2)(x) &= f_2(xg_1) \\ &= (g_2 f)(xg_1) = f(xg_1 g_2) = [(g_1 g_2)f](x),\end{aligned}$$

so $g_1(g_2 f) = (g_1 g_2)f$. Also, clearly $1f = f$ for all $f \in W$, where 1 is the identity of G. This action of G on W may be extended to an action of FG on W by $(\sum \alpha_g g)f = \sum \alpha_g (gf)$. In this way W becomes an FG-module.

Definition. The FG-module W constructed above is called the *induced module* of V and the representation of G it affords is called the *induced representation* of φ. If ψ is the character of φ then the character of the induced representation is called the *induced character* and is denoted by $\mathrm{Ind}_H^G(\psi)$.

When it is necessary to specify the groups involved in the discussion of induced representations we shall say that W is the FG-module (or representation) induced from the FH-module V (for characters, the groups involved are evident from the notation $\mathrm{Ind}_H^G(\psi)$).

Proposition 12. If V is an FH-module which affords a representation of H of degree n then the induced representation is a representation of G of degree $n \cdot |G : H|$. If ψ is the character afforded by V then the value of the induced character on an element $g \in G$ is given by the formula

$$\mathrm{Ind}_H^G(\psi)(g) = \sum_{i=1}^{m} \psi(g_i g g_i^{-1})$$

where $g_1, g_2, \ldots, g_m$ are any set of representatives for the right cosets of H in G and $\psi(g_i g g_i^{-1})$ is defined to be 0 if $g_i g g_i^{-1} \notin H$.

Proof: Let W be the induced module of V described above. Let $g_1, \ldots, g_m$ be a set of representatives for the right cosets of H in G. Each $f \in W$ is uniquely determined by its values $f(g_1), \ldots, f(g_m)$. This is because any $g \in G$ lies in a unique right coset Hg_j, so $g = hg_j$ for some $h \in H$ and $f(g) = f(hg_j) = hf(g_j)$. Let

$$W_j = \{f \in W \mid f(hg_k) = 0 \text{ for all } h \in H \text{ and all } k \neq j\}.$$

It is easy to see that

$$W = W_1 \oplus W_2 \oplus \cdots \oplus W_m.$$

We show that each W_j has the same dimension as V, which will prove the first assertion of the proposition. Let $\dim V = n$ and let $e_1, e_2, \ldots, e_n$ be a basis of V. Fix j and let $f_{j1}, \ldots, f_{jn}$ be the elements of W_j defined by

$$f_{jk}(g_i) = \begin{cases} e_k & \text{if } i = j \quad \text{for } 1 \leq k \leq n \\ 0 & \text{if } i \neq j \quad \text{for } 1 \leq k \leq n \end{cases}$$

(so $f_{jk}(hg_j) = hf_{jk}(g_j) = he_k$ and f_{jk} is zero on Hg_i for all $i \neq j$). Since $e_1, \ldots, e_n$ are linearly independent, so are $f_{j1}, \ldots, f_{jn}$ and these are easily seen to span W_j. Thus a basis of W_j has cardinality n so the dimension of W_j equals the dimension of V as claimed.

Let g be an arbitrary element of G. We compute the matrix of g on W with respect to the basis $\{f_{jk} \mid 1 \leq j \leq m,\ 1 \leq k \leq n\}$. Since f_{jk} is zero on every right coset except Hg_j we have

$$(gf_{jk})(g_i) = f_{jk}(g_i g) = \begin{cases} (g_i g g_j^{-1}) f_{jk}(g_j) & \text{if } g_i g = hg_j \in Hg_j \\ 0 & \text{if } g_i g \notin Hg_j \end{cases}$$

for all $k \in \{1, 2, \ldots, n\}$. Thus in its action on W each g permutes the direct summands $W_1, \ldots, W_m$ by sending the summand W_j to W_i if $g_i g \in Hg_j$. The matrix for g with respect to this basis is an $nm \times nm$ block matrix, that is, an $m \times m$ matrix each of whose entries is an $n \times n$ matrix. The ith row and j^{th} column of this $m \times m$ matrix is an $n \times n$ zero matrix if $g_i g \notin Hg_j$; otherwise, if $g_i g = hg_j \in Hg_j$ then this entry is the $n \times n$ matrix of h acting on V with respect to the basis $e_1, \ldots, e_n$. In other words, if $T : H \to GL_n(F)$ is the matrix representation afforded by the action of H on V with respect to the basis $e_1, \ldots, e_n$, then the matrix representation of the induced representation $T^* : G \to GL_{nm}(F)$ may be described by $T^*(g) = (T(g_i g g_j^{-1}))_{i,j}$, where the $n \times n$ matrix $T(g_i g g_j^{-1})$ is defined to be the zero matrix if $g_i g g_j^{-1} \notin H$ (an explicit example of some matrices for an induced representation is given below).

It is now easy to read off the character of g acting on W from the matrix of g computed above. The blocks $T(g_i g g_i^{-1})$ down the diagonal of $T^*(g)$ are zero except when $g_i g g_i^{-1} \in H$. Thus the trace of the block matrix $T^*(g)$ is the sum of the traces of the matrices $T(g_i g g_i^{-1})$ for which $g_i g g_i^{-1} \in H$. Since the trace of $T(g_i g g_i^{-1})$ is $\psi(g_i g g_i^{-1})$ the proof is complete.

Remark: Since the character ψ of H is constant on the conjugacy classes of H we have $\psi(g) = \psi(hgh^{-1})$ for all $h \in H$. Furthermore, as h runs over all elements of H, hx runs over all elements of the coset Hx. Thus, the formula for the induced character may also be written

$$\operatorname{Ind}_H^G(\psi)(g) = \frac{1}{|H|} \sum_{x \in G} \psi(xgx^{-1})$$

where the elements x in each fixed coset give the same character value $|H|$ times (which accounts for the factor of $\frac{1}{|H|}$) and again, $\psi(xgx^{-1}) = 0$ if $xgx^{-1} \notin H$.

Corollary 13. In the notation of Proposition 12, $\operatorname{Ind}_H^G(\psi)(g) = 0$ if g is not conjugate in G to some element of H. In particular, if H is a normal subgroup of G then $\operatorname{Ind}_H^G(\psi)$ is zero on all elements of $G - H$.

Proof: If $g_i g g_i^{-1} \notin H$ for all coset representatives g_i then each term in the sum for $\operatorname{Ind}_H^G(\psi)(g)$ is zero. In particular, if g is not in the normal subgroup H then neither is any conjugate of g, so $\operatorname{Ind}_H^G(\psi)$ is zero on g.

Examples

(1) Let $G = D_{12} = \langle r, s \mid r^6 = s^2 = 1,\ rs = sr^{-1}\rangle$ be the dihedral group of order 12 and let $H = \langle s, r^3\rangle$ so that H is isomorphic to the Klein 4-group and $|G : H| = 3$. Following the notation and proof of Proposition 12 we exhibit the matrices for r and s of the induced representation of a specific representation of H. Let the representation of H on a 2-dimensional vector space over $\mathbb{Q}$ with respect to some basis e_1, e_2 be given by

$$s \mapsto \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = A, \qquad r^3 \mapsto \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = B, \qquad sr^3 \mapsto \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = C,$$

so $n = 2$, $m = 3$ and the induced representation has degree $nm = 6$. Fix representatives $1, r, r^2$ for the right cosets of H in G. The action of r by right multiplication on these cosets and on their representatives is given by

$$\begin{array}{lll} r : H1 \mapsto Hr & r : Hr \mapsto Hr^2 & r : Hr^2 \mapsto H1 \\ \quad 1 \mapsto r & \quad r \mapsto r^2 & \quad r^2 \mapsto r^3 \end{array}$$

and the action of s is given by

$$\begin{array}{lll} s : H1 \mapsto H1 & s : Hr \mapsto Hr^2 & s : Hr^2 \mapsto Hr \\ \quad 1 \mapsto s & \quad r \mapsto rs = (sr^3)r^2 & \quad r^2 \mapsto r^2 s = (sr^3)r. \end{array}$$

Thus r and s permute the three subspaces W_1, W_2 and W_3 described in the proof of Proposition 12 by

$$\begin{array}{lll} r : W_2 \mapsto W_1 & r : W_3 \mapsto W_2 & r : W_1 \mapsto W_3 \\ s : W_1 \mapsto W_1 & s : W_3 \mapsto W_2 & s : W_2 \mapsto W_3. \end{array}$$

From these calculations the 6×6 matrices for r and s may be described as 3×3 block matrices whose entries are 2×2 matrices. In the 3×3 block matrix for s (or for r) there is a nonzero block in row i and column j whenever s (or r, respectively) sends W_j to W_i. Thus the nonzero blocks for s appear in the 1,1 position, the 2,3 position and the 3,2 position of the 3×3 block matrix. The 2×2 matrices which appear in these positions are computed for s (and analogously for r) as follows: if s maps W_j to W_i then $g_i s = hg_j$ for some $h \in H$, where g_1, g_2, g_3 are the fixed coset representatives $1, r, r^2$ of H in G. Then the 2×2 matrix for this element h in the original representation of H (i.e. either A, B, C or the 2×2 identity matrix) appears in position i, j. Thus, for example, s maps W_3 to W_2 and $rs = (sr^3)r^2$, and so the matrix C appears in position 2,3 for s. The matrices for r and s in this induced representation are therefore:

$$r \mapsto \begin{pmatrix} 0 & I & 0 \\ 0 & 0 & I \\ B & 0 & 0 \end{pmatrix} \qquad s \mapsto \begin{pmatrix} A & 0 & 0 \\ 0 & 0 & C \\ 0 & C & 0 \end{pmatrix}$$

where the 2×2 matrices A, B and C are given above, I is the 2×2 identity matrix and 0 denotes the 2×2 zero matrix.

(2) If H is any subgroup of G and ψ_1 is the principal character of H, then $\mathrm{Ind}_H^G(\psi_1)(g)$ counts 1 for each coset representative g_i such that $g_i g g_i^{-1} \in H$. Since $g_i g g_i^{-1} \in H$ if and only if g fixes the right coset Hg_i under right multiplication (or fixes the left coset $g_i^{-1}H$ under left multiplication), $\mathrm{Ind}_H^G(\psi_1)(g)$ is the number of points fixed

by g in the permutation representation of g on the (left or right) cosets of H. Thus by Example 3 of Section 15.4 we see that: *if ψ_1 is the principal character of H then $Ind_H^G(\psi_1)$ is the permutation character on the left cosets of H in G.* In the special case when $H = 1$, this implies *if χ_1 is the principal character of the trivial subgroup $H = 1$ then $Ind_1^G(\chi_1)$ is the regular character of G.* This also shows that an induced character is not, in general, irreducible even if the character from which it is induced is irreducible.

(3) Let $G = S_3$ and let ψ be a nonprincipal linear character of $A_3 = \langle x \rangle$, so that $\psi(x) = \zeta$, for some primitive cube root of unity ζ (the character tables of $A_3 = Z_3$ and S_3 appear in Section 1). Let $\Psi = \mathrm{Ind}_{A_3}^{S_3}(\psi)$. Thus Ψ has degree $1 \cdot |S_3 : A_3| = 2$ and, by the corollary, Ψ is zero on all transpositions. If y is any transposition then $1, y$ is a set of right coset representatives of A_3 in S_3 and $yxy^{-1} = x^2$. Thus $\Psi(x) = \psi(x) + \psi(x^2) = \zeta + \zeta^2 = -1$. This shows that if ψ is either of the two nonprincipal irreducible characters of A_3 then the induced character of ψ is the (unique) irreducible character of S_3 of degree 2. In particular, different characters of a subgroup may induce the same character of the whole group.

(4) Let $G = D_8$ have its usual generators and relations and let $H = \langle s \rangle$. Let ψ be the nonprincipal irreducible character of H and let $\Psi = \mathrm{Ind}_H^G(\psi)$. Pick right coset representatives $1, r, r^2, r^3$ for H. By Proposition 12, $\Psi(1) = 4$. Since $\psi(s) = -1$, one computes directly that $\Psi(s) = -2$. By Corollary 13 we obtain $\Psi(r) = \Psi(r^2) = \Psi(sr) = 0$. In the notation of the character table of D_8 in Section 1, by the orthogonality relations we obtain $\Psi = \chi_2 + \chi_4 + \chi_5$ (which may be checked by inspection).

For the remainder of this section the field F is taken to be the complex numbers: $F = \mathbb{C}$.

Before concluding with an application of induced characters to simple groups we compute the characters of an important class of groups.

Definition. A finite group G is called a *Frobenius group* with *Frobenius kernel* Q if Q is a normal subgroup of G and $C_G(x) \leq Q$ for all nonidentity elements x of Q.

In view of the application to simple groups mentioned at the beginning of this section we shall restrict attention to Frobenius groups G of order $q^a p$, where p and q are distinct primes, such that the Frobenius kernel Q is an elementary abelian q-group of order q^a and the cyclic group G/Q acts irreducibly by conjugation on Q. In other words, we shall assume Q is a direct product of cyclic groups of order q and the only normal subgroups of G that are contained in Q are 1 and Q, i.e. Q is a minimal normal subgroup of G. For example, A_4 is a Frobenius group of this type with Frobenius kernel V_4, its Sylow 2-subgroup. Also, if p and q are distinct primes with $p < q$ and G is a non-abelian group of order pq (one always exists if $p \mid q-1$) then G is a Frobenius group whose Frobenius kernel is its Sylow q-subgroup (which is normal by Sylow's Theorem). We essentially determine the character table of these Frobenius groups. Analogous results on more general Frobenius groups appear in the exercises.

Proposition 14. Let G be a Frobenius group of order $q^a p$, where p and q are distinct primes, such that the Frobenius kernel Q is an elementary abelian q-group of order q^a and the cyclic group G/Q acts irreducibly by conjugation on Q. Then the following hold:

(1) $G = QP$ where P is a Sylow p-subgroup of G. Every nonidentity element of G has order p or q. Every element of order p is conjugate to an element of P and every element of order q belongs to Q. The nonidentity elements of P represent the $p-1$ distinct conjugacy classes of elements of order p and each of these classes has size q^a. There are $(q^a-1)/p$ distinct conjugacy classes of elements of order q and each of these classes has size p.

(2) $G' = Q$ so the number of degree 1 characters of G is p and every degree 1 character contains Q in its kernel.

(3) If ψ is any nonprincipal irreducible character of Q, then $\mathrm{Ind}_Q^G(\psi)$ is an irreducible character of G. Moreover, every irreducible character of G of degree > 1 is equal to $\mathrm{Ind}_Q^G(\psi)$ for some nonprincipal irreducible character ψ of Q. Every irreducible character of G has degree either 1 or p and the number of irreducible characters of degree p is $(q^a-1)/p$.

Proof: Note that QP equals G by order consideration. By definition of a Frobenius group and because Q is abelian, $C_G(h) = Q$ for every nonidentity element h of Q. If x were an element of order pq, then x^p would be an element of order q, hence would lie in the unique Sylow q-subgroup Q of G. But then x would commute with x^p and so x would belong to $C_G(x^p) = Q$, a contradiction. Thus G has no elements of order pq. By Sylow's Theorem every element of order p is conjugate to an element of P and every element of order q lies in Q. No two distinct elements of P are conjugate in G because if $g^{-1}xg = y$ for some $x, y \in P$ then $\overline{g^{-1}xg} = \overline{y}$ in the abelian group $\overline{G} = G/Q$ and so $\overline{x} = \overline{y}$. Then $x = y$ because $\overline{P} \cong P$. Thus there are exactly $p-1$ conjugacy classes of elements of order p and these are represented by the nonidentity elements of P. If x is a nonidentity element of P, then $C_G(x) = P$ and so the conjugacy class of x consists of $|G : P| = q^a$ elements. Finally, if h is a nonidentity element of Q, then $C_G(h) = Q$ and the conjugacy class of h is $\{h, h^x, \ldots, h^{x^{p-1}}\}$, where $P = \langle x \rangle$. This proves all parts of (1).

Since G/Q is abelian, $G' \leq Q$. Since G is non-abelian and Q is, by hypothesis, a minimal normal subgroup of G we must have $G' = Q$. Part (2) now follows from Corollary 15 in Section 15.3.

Let ψ be a nonprincipal irreducible character of Q and let $\Psi = \mathrm{Ind}_Q^G(\psi)$. We use the orthogonality relations to show that Ψ is irreducible. Let $1, x, \ldots, x^{p-1}$ be coset representatives for Q in G. By Corollary 13, Ψ is zero on $G - Q$ so

$$\begin{aligned}
||\Psi||^2 &= \frac{1}{|G|} \sum_{h \in Q} \Psi(h)\overline{\Psi(h)} \\
&= \frac{1}{|G|} \sum_{h \in Q} \sum_{i=0}^{p-1} \psi(x^i h x^{-i})\overline{\psi(x^i h x^{-i})} \\
&= \frac{p}{|G|} \sum_{h \in Q} \psi(h)\overline{\psi(h)} \\
&= \frac{p|Q|}{|G|} = 1.
\end{aligned}$$

where the second line follows from the definition of the induced character Ψ, the third line follows because each element of Q appears exactly p times in the sum in the second line, and the last line follows from the first orthogonality relation in Q because ψ is an irreducible character of Q. This proves Ψ is an irreducible character of G.

We prove that every irreducible character of G of degree > 1 is the induced character of some nonprincipal degree 1 character of Q by counting the number of distinct irreducible characters of G obtained this way. By parts (1) and (2) the number of irreducible characters of G (= the number of conjugacy classes) is $p+(q^a-1)/p$ and the number of degree 1 characters is p. Thus the number of irreducible characters of G of degree > 1 is $(q^a-1)/p$. The group P acts on the set $\mathcal{C}$ of nonprincipal irreducible characters of Q as follows: for each $\psi \in \mathcal{C}$ and each $x \in P$ let ψ^x be defined by

$$\psi^x(h) = \psi(xhx^{-1}) \qquad \text{for all } h \in Q.$$

Since ψ is a nontrivial homomorphism from Q into $\mathbb{C}^\times$ (recall that all irreducible characters of the abelian group Q have degree 1) it follows easily that ψ^x is also a homomorphism. Thus $\psi^x \in \mathcal{C}$ and so P permutes the elements of $\mathcal{C}$. Now let x be a generator for the cyclic group P. Then $1, x, \ldots, x^{p-1}$ are representatives for the right cosets of Q in G. By Proposition 12 applied with this set of coset representatives we see that if $\psi \in \mathcal{C}$ then the value of $\mathrm{Ind}_Q^G(\psi)$ on any element h of Q is $\psi(h) + \psi^x(h) + \cdots + \psi^{x^{p-1}}(h)$. Thus when the induced character $\mathrm{Ind}_Q^G(\psi)$ is restricted to Q it decomposes into irreducible characters of Q as

$$\mathrm{Ind}_Q^G(\psi)|_Q = \psi + \psi^x + \cdots + \psi^{x^{p-1}}.$$

If ψ_1 and ψ_2 are in different orbits of the action of P on $\mathcal{C}$ then the induced characters $\mathrm{Ind}_Q^G(\psi_1)$ and $\mathrm{Ind}_Q^G(\psi_2)$ restrict to distinct characters of Q (they have no irreducible constituents in common). Thus characters induced from elements of distinct orbits of P on $\mathcal{C}$ are distinct irreducible characters of G. The abelian group Q has $q^a - 1$ nonprincipal irreducible characters (i.e. $|\mathcal{C}| = q^a - 1$) and $|P| = p$ so there are at least $(q^a - 1)/p$ orbits of P on $\mathcal{C}$ and hence at least this number of distinct irreducible characters of G of degree p. Since G has exactly $(q^a - 1)/p$ irreducible characters of degree > 1, every irreducible character of G of degree > 1 must have degree p and must be an induced character from some element of $\mathcal{C}$. The proof is complete.

For the final example we shall require two properties of induced characters. These properties are listed in the next proposition and the proofs, which follow easily from the formula for induced characters, are left as exercises.

Proposition 15. Let G be a group, let H be a subgroup of G and let ψ and ψ' be characters of H.

(1) (Induction of characters is additive) $\mathrm{Ind}_H^G(\psi + \psi') = \mathrm{Ind}_H^G(\psi) + \mathrm{Ind}_H^G(\psi')$.

(2) (Induction of characters is transitive) If $H \le K \le G$ then

$$\mathrm{Ind}_K^G(\mathrm{Ind}_H^K(\psi)) = \mathrm{Ind}_H^G(\psi).$$

It follows from part (1) of Proposition 15 that if $\sum_{i=1}^{s} n_i\psi_i$ is any integral linear combination of characters of H with $n_i \geq 0$ for all i then

$$\operatorname{Ind}_H^G\left(\sum_{i=1}^{s} n_i\psi_i\right) = \sum_{i=1}^{s} n_i \operatorname{Ind}_H^G(\psi_i). \qquad (*)$$

A class function of H of the form $\sum_{i=1}^{s} n_i\psi_i$, where the coefficients are any integers (not necessarily nonnegative) is called a *generalized character* or *virtual character* of H. For a generalized character of H we define its induced generalized character of G by equation $(*)$, allowing now negative coefficients n_i as well. In this way the function Ind_H^G becomes a group homomorphism from the additive group of generalized characters of H to the additive group of generalized characters of G (which maps characters to characters). This implies that the formula for induced characters in Proposition 12 holds also if ψ is a generalized character of H.

We now conclude with a proof of the following result:

there are no simple groups of order $3^3 \cdot 7 \cdot 13 \cdot 409$.

As mentioned at the beginning of this section, simple groups of this order were discussed at the end of Section 6.2 in the context of the existence problem for simple groups. It is possible to prove that there are no simple groups of this order by arguments involving a permutation representation of degree 819 (cf. the exercises in Section 6.2). We include a character-theoretic proof of this since the methods illustrate some important ideas in the theory of finite groups. The approach is based on M. Suzuki's seminal paper — The nonexistence of a certain type of simple group of odd order, *Jour. of the London Mathematical Society*, 1957 — which treats much more general groups. Because we are dealing with a specific group order, our arguments are simpler and numerically more explicit, yet they retain some of the key ideas of Suzuki's work. Moreover, Suzuki's paper and its successor, Finite groups in which the centralizer of any non-identity element is nilpotent by W. Feit, M. Hall and J. Thompson, *Mathematisches Zeitschrift*, 1960, are prototypes for the lengthy and difficult Feit-Thompson Theorem (cf. Section 3.4). Our discussion also conveys some of the flavor of these fundamental papers. In particular, each of these papers follows the basic development in which the structure and embedding of the Sylow subgroups is first determined and then character theory (with heavy reliance on induced characters) is applied.

For the remainder of this section we assume G is a simple group of order $3^3 \cdot 7 \cdot 13 \cdot 409$. We list some properties of G which may be verified using the methods stemming from Sylow's Theorem discussed in Section 6.2. The details are left as exercises.

(1) Let $q_1 = 3$, let Q_1 be a Sylow 3-subgroup of G and let $N_1 = N_G(Q_1)$. Then Q_1 is an elementary abelian 3-group of order 3^3 and N_1 is a Frobenius group of order $3^3 \cdot 13$ with Frobenius kernel Q_1 and with N_1/Q_1 acting irreducibly by conjugation on Q_1.

(2) Let $q_2 = 7$, let Q_2 be a Sylow 7-subgroup of G and let $N_2 = N_G(Q_2)$. Then Q_2 is cyclic of order 7 and N_2 is the non-abelian group of order $7 \cdot 3$ (so N_2 is a Frobenius group with Frobenius kernel Q_2).

(3) Let $q_3 = 13$, let Q_3 be a Sylow 11-subgroup of G and let $N_3 = N_G(Q_3)$. Then

Q_3 is cyclic of order 13 and N_3 is the non-abelian group of order $13 \cdot 3$ (so N_3 is a Frobenius group with Frobenius kernel Q_3).

(4) Let $q_4 = 409$, let Q_4 be a Sylow 409-subgroup of G and let $N_4 = N_G(Q_4)$. Then Q_4 is cyclic of order 409 and N_4 is the non-abelian group of order $409 \cdot 3$ (so N_4 is a Frobenius group with Frobenius kernel Q_4).

(5) Every nonidentity element of G has prime order and $Q_i \cap Q_i^g = 1$ for every $g \in G - N_i$, for each $i = 1, 2, 3, 4$. The nonidentity conjugacy classes of G are:

(a) 2 classes of elements of order 3 (each of these classes has size $7 \cdot 13 \cdot 409$)

(b) 2 classes of elements of order 7 (each of these classes has size $3^3 \cdot 13 \cdot 409$)

(c) 4 classes of elements of order 13 (each of these classes has size $3^3 \cdot 7 \cdot 409$)

(d) 136 classes of elements of order 409 (each of these classes has size $3^3 \cdot 7 \cdot 13$),

and so there are 145 conjugacy classes in G.

Since each of the groups N_i is a Frobenius group satisfying the hypothesis of Proposition 14, the number of characters of N_i of degree > 1 may be read off from that proposition:

(i) N_1 has 2 irreducible characters of degree 13

(ii) N_2 has 2 irreducible characters of degree 3

(iii) N_3 has 4 irreducible characters of degree 3

(iv) N_4 has 136 irreducible characters of degree 3.

From now on, to simplify notation, for any subgroup H of G and any generalized character μ of H let

$$\mu^* = \mathrm{Ind}_H^G(\mu)$$

so a star will always denote induction from a subgroup to the whole group G and the subgroup will be clear from the context.

The following lemma is a key point in the proof. It shows how the vanishing of induced characters described in Corollary 13 (together with the *trivial intersection* property of the Sylow subgroups Q_i, namely the fact that $Q_i \cap Q_i^g = 1$ for all $g \in G - N_G(Q_i)$) may be used to relate inner products of certain generalized characters to the inner products of their induced generalized characters. For these computations it is important that the generalized characters are zero on the identity (which explains why we are considering *differences* of characters of the same degree).

Lemma 16. For any $i \in \{1, 2, 3, 4\}$ let $q = q_i$, let $Q = Q_i$, let $N = N_i$ and let $p = |N : Q|$. Let $\psi_1, \ldots, \psi_4$ be any irreducible characters of N of degree p (not necessarily distinct) and let $\alpha = \psi_1 - \psi_2$ and $\beta = \psi_3 - \psi_4$. Then α and β are generalized characters of N which are zero on every element of N of order not equal to q. Furthermore, α^* and β^* are generalized characters of G which are zero on every element of G of order not equal to q and

$$(\alpha^*, \beta^*)_G = (\alpha, \beta)_N$$

(where $(\ , \ )_H$ denotes the usual Hermitian product of class functions computed in the group H). In other words, induction from N to G is an inner product preserving map on such generalized characters α, β of N.

Proof: By Proposition 14, there are nonprincipal characters $\lambda_1, \ldots, \lambda_4$ of Q of degree 1 such that $\psi_j = \mathrm{Ind}_Q^N(\lambda_j)$ for $j = 1, \ldots, 4$. By Corollary 13 therefore, each ψ_j vanishes on $N - Q$, hence so do α and β. Note that since $\psi_j(1) = p$ for all j we have $\alpha(1) = \beta(1) = 0$. By the transitivity of induction, $\psi_j^* = \mathrm{Ind}_N^G(\psi_j) = \mathrm{Ind}_Q^G(\lambda_j)$ for all j. Again by Corollary 13 applied to the latter induced character we see that ψ_j^* vanishes on all elements not conjugate in G to some element of Q, hence so do both α^* and β^*. Since the induced characters ψ_j^* all have degree $|G : Q|$, the generalized characters α^* and β^* are zero on the identity. Thus α^* and β^* vanish on all elements of G which are not of order q. Finally, if $g_1, \ldots, g_m$ are representatives for the right cosets of N in G with $g_1 = 1$, then because $Q \cap Q^{g_k} = 1$ for all $k > 1$ (by (5) above), it follows immediately from the formula for induced (generalized) characters that $\alpha^*(x) = \alpha(x)$ and $\beta^*(x) = \beta(x)$ for all nonidentity elements $x \in Q$ (i.e. for all elements $x \in N$ of order q). Furthermore, by Sylow's Theorem every element of G of order q lies in a conjugate of Q, hence the collection of G-conjugates of the set $Q - \{1\}$ partition the elements of order q in G into $|G : N|$ disjoint subsets. Since α^* and β^* are class functions on G, the sum of $\alpha^*(x)\overline{\beta^*(x)}$ as x runs over any of these subsets is the same. These facts imply

$$\begin{aligned}
(\alpha^*, \beta^*)_G &= \frac{1}{|G|} \sum_{x \in G} \alpha^*(x)\overline{\beta^*(x)} \\
&= \frac{1}{|G|} \sum_{\substack{x \in G \\ |x| = q}} \alpha^*(x)\overline{\beta^*(x)} \\
&= \frac{1}{|G|} \sum_{\substack{x \in N \\ |x| = q}} |G : N| \alpha^*(x)\overline{\beta^*(x)} \\
&= \frac{1}{|N|} \sum_{x \in N} \alpha(x)\overline{\beta(x)} = (\alpha, \beta)_N.
\end{aligned}$$

This completes the proof.

The next lemma sets up a correspondence between the irreducible characters of N_i of degree > 1 and some nonprincipal irreducible characters of G.

Lemma 17. For any $i \in \{1, 2, 3, 4\}$ let $q = q_i$, let $Q = Q_i$, let $N = N_i$ and let $p = |N : Q|$. Let $\psi_1, \ldots, \psi_k$ be the distinct irreducible characters of N of degree p. Then there are distinct irreducible characters $\chi_1, \ldots, \chi_k$ of G, all of which have the same degree, and a fixed sign $\epsilon = \pm 1$ such that $\psi_1^* - \psi_j^* = \epsilon(\chi_1 - \chi_j)$ for all $j = 2, 3, \ldots, k$.

Proof: Let $\alpha_j = \psi_1 - \psi_j$ for $j = 2, 3, \ldots, k$ so α_j satisfies the hypothesis of Lemma 16. Since $\psi_1 \neq \psi_j$, by Lemma 16

$$2 = ||\alpha_j||^2 = (\alpha_j, \alpha_j)_N = (\alpha_j^*, \alpha_j^*)_G = ||\alpha_j^*||^2$$

for all j. Thus α_j^* must have two distinct irreducible characters of G as its irreducible constituents. Since $\alpha_j^*(1) = 0$ it must be a difference of two distinct irreducible char-

acters, both of which have the same degree. In particular, the lemma holds if $k = 2$ (which is the case for $q = 3$ and $q = 7$). Assume therefore that $k > 2$ and write

$$\alpha_2^* = \psi_1^* - \psi_2^* = \epsilon(\chi - \chi')$$
$$\alpha_3^* = \psi_1^* - \psi_3^* = \epsilon'(\theta - \theta')$$

for some irreducible characters $\chi, \chi', \theta, \theta'$ of G and some signs ϵ, ϵ'. As proved above, $\chi \neq \chi'$ and $\theta \neq \theta'$. Interchanging θ and θ' if necessary, we may assume $\epsilon = \epsilon'$. Thus

$$\alpha_3^* - \alpha_2^* = \psi_2^* - \psi_3^* = \epsilon(\theta - \theta' - \chi + \chi').$$

By Lemma 16, $\psi_2^* - \psi_3^* = (\psi_2 - \psi_3)^*$ also has exactly two distinct irreducible constituents, hence either $\theta = \chi$ or $\theta' = \chi'$. Replacing ϵ by $-\epsilon$ if necessary we may assume that $\theta = \chi$ so that now we have

$$\alpha_2^* = \psi_1^* - \psi_2^* = \epsilon(\chi - \chi')$$
$$\alpha_3^* = \psi_1^* - \psi_3^* = \epsilon(\chi - \theta')$$

where χ, χ' and θ are distinct irreducible characters of G and the sign ϵ is determined. Label $\chi = \chi_1$, $\chi' = \chi_2$ and $\theta = \chi_3$. Now one similarly checks that for each $j \geq 3$ there is an irreducible character χ_j of G such that

$$\alpha_j^* = \psi_1^* - \psi_j^* = \epsilon(\chi_1 - \chi_j)$$

and $\chi_1, \ldots, \chi_k$ are distinct. Since all χ_j's have the same degree as χ_1, the proof is complete.

We remark that it need not be the case that $\chi_j = \psi_j^*$ for any j, but only that the differences of irreducible characters of N induce to differences of irreducible characters of G.

The irreducible characters χ_j of G obtained via Lemma 17 are called *exceptional characters* associated to Q.

Lemma 18. The exceptional characters associated to Q_i are all distinct from the exceptional characters associated to Q_j for i and j distinct elements of $\{1, 2, 3, 4\}$.

Proof: Let χ be an exceptional character associated to Q_i and let θ be an exceptional character associated to Q_j. By construction, there are distinct irreducible characters ψ and ψ' of Q_i such that $\psi^* - \psi'^* = \chi - \chi'$ and there are distinct irreducible characters λ and λ' of Q_j such that $\lambda^* - \lambda'^* = \theta - \theta'$. Let $\alpha = \psi - \psi'$ and let $\beta = \lambda - \lambda'$. By Lemma 16, α^* is zero on all elements of G whose order is not equal to q_i (including the identity) and β^* is zero on all elements of G whose order is not equal to q_j. Thus clearly $(\alpha^*, \beta^*) = 0$. It follows easily that the two irreducible constituents of α^* are pairwise orthogonal to those of β^* as well. This establishes the lemma.

It is now easy to show that such a simple group G does not exist. By Lemma 17 and properties (i) to (iv) of G we can count the number of exceptional characters:

(i) there are 2 exceptional characters associated to Q_1

(ii) there are 2 exceptional characters associated to Q_2

(iii) there are 4 exceptional characters associated to Q_3

(iv) there are 136 exceptional characters associated to Q_4 .

Denote the common degree of the exceptional characters associated to Q_i by d_i for $i = 1, \ldots, 4$. By Lemma 18, the exceptional characters account for 144 nonprincipal irreducible characters of G hence these, together with the principal character, are all the irreducible characters of G (the number of conjugacy classes of G is 145). The sum of the squares of the degrees of the irreducible characters is the order of G:

$$1 + 2d_1^2 + 2d_2^2 + 4d_3^2 + 136d_4^2 = 1004913.$$

Simplifying this, we obtain

$$d_1^2 + d_2^2 + 2d_3^2 + 68d_4^2 = 502456. \tag{16.4}$$

Finally, since each nonprincipal irreducible representation of the simple group G is faithful and since the smallest degree of a faithful representation of N_1 is 13, each $d_i \geq 13$. Since $d_4 < \sqrt{50245668} < 86$ and d_4 divides $|G|$, it follows that

$$d_4 \in \{13, 21, 27, 39, 63\}.$$

Furthermore, each $d_i \mid |G|$ by Corollary 6 and so there are a small number of possibilities for each d_i. One now checks that equation (16.4) has no solution (this is particularly easy to do by computer). This contradiction completes the proof.

EXERCISES

Throughout the exercises all representations are over the complex numbers.

1. Let $G = S_3$, let $H = A_3$ and let V be the 3-dimensional $\mathbb{C}H$-module which affords the natural permutation representation of A_3. More explicitly, let V have basis e_1, e_2, e_3 and let $\sigma \in A_3$ act on V by $\sigma e_i = e_{\sigma(i)}$. Let 1 and (1 2) be coset representatives for the right cosets of A_3 in S_3 and write out the explicit matrices described in the proof of Proposition 12 for the action of S_3 on the induced module W, for each of the elements of S_3.
2. In each of parts (a) to (f) a character ψ of a subgroup H of a particular group G is specified. Compute the values of the induced character $\mathrm{Ind}_H^G(\psi)$ on all the conjugacy classes of G and use the character tables in Section 1 to write $\mathrm{Ind}_H^G(\psi)$ as a sum of irreducible characters:
   (a) ψ is the unique nonprincipal degree 1 character of the subgroup $\langle(1\ 2)\rangle$ of S_3
   (b) ψ is the degree 1 character of the subgroup $\langle r\rangle$ of D_8 defined by $\psi(r) = i$, where $i \in \mathbb{C}$ is the square root of -1
   (c) ψ is the degree 1 character of the subgroup $\langle r\rangle$ of D_8 defined by $\psi(r) = -1$
   (d) ψ is any of the three nonprincipal degree 1 characters of the subgroup $V_4 = \langle(1\ 2), (3\ 4)\rangle$ of S_4
   (e) $\psi = \chi_4$ is the first of the two characters of degree 3 in the character table of $H = S_4$ in Section 1 and H is a subgroup of $G = S_5$
   (f) ψ is any of the three nonprincipal degree 1 characters of the subgroup $V_4 = \langle(1\ 2), (3\ 4)\rangle$ of S_5.
3. Use Proposition 14 to explicitly write out the character table of each of the following groups:
   (a) the dihedral group of order 10
   (b) the non-abelian group of order 57

(c) the non-abelian group of order 56 which has a normal, elementary abelian Sylow 2-subgroup.

4. Let H be a subgroup of G, let φ be a representation of H and suppose that N is a normal subgroup of G with $N \leq H$ and N contained in the kernel of φ. Prove that N is also contained in the kernel of the induced representation of φ.

5. Let N be a normal subgroup of G and let ψ_1 be the principal character of N. Let $\Psi = \mathrm{Ind}_N^G(\psi_1)$ so that by the preceding exercise we may consider Ψ as the character of a representation of G/N. Prove that Ψ is the character of the regular representation of G/N.

6. Let Z be any subgroup of the center of G, let $|G : Z| = m$ and let ψ be a character of Z. Prove that

$$\mathrm{Ind}_Z^G(\psi)(g) = \begin{cases} m\psi(g) & \text{if } g \in Z \\ 0 & \text{if } g \notin Z. \end{cases}$$

7. Let G be a Frobenius group with Frobenius kernel Q. Assume that both Q and G/Q are abelian but G is not abelian (i.e. $G \neq Q$). Let $|Q| = n$ and $|G : Q| = m$.
   (a) Prove that G/Q is cyclic and show that $G = QC$ for some cyclic subgroup C of G with $C \cap Q = 1$ (i.e. G is a semidirect product of Q and C and $|C| = m$). [Let q be a prime divisor of n and let G/Q act by conjugation on the elementary abelian q-group $\{h \in Q \mid h^q = 1\}$. Apply Exercise 14(e) of Section 15.1 and the definition of a Frobenius group to an irreducible constituent of this $\mathbb{F}_q G/Q$-module.]
   (b) Prove that n and m are relatively prime. [If a prime p divides both the order and index of Q, let P be a Sylow p-subgroup of G. Then $P \cap Q \trianglelefteq P$ and $P \cap Q$ is a Sylow p-subgroup of Q. Consider the centralizer in G of the subgroup $Z(P) \cap Q$ (this intersection is nontrivial by Theorem 1 of Section 6.1).]
   (c) Show that G has no elements of order qp, where q is any nontrivial divisor of n and p is any nontrivial divisor of m. [Argue as in Proposition 14.]
   (d) Prove that the number of nonidentity conjugacy classes of G contained in Q is $(n-1)/m$ and that each of these classes has size m. [Argue as in Proposition 14.]
   (e) Prove that no two distinct elements of C are conjugate in G. Deduce that the nonidentity elements of C are representatives for $m-1$ distinct conjugacy classes of G and that each of these classes has size n. Deduce then that every element of $G - Q$ is conjugate to some element of C and that G has $m + (n-1)/m$ conjugacy classes.
   (f) Prove that $G' = Q$ and deduce that G has m distinct characters of degree 1. [To show $Q \leq G'$ let $C = \langle x \rangle$ and argue that the map $h \mapsto [h, x] = x^{-1}h^{-1}xh$ is a homomorphism from Q to Q whose kernel is trivial, hence this map is surjective.]
   (g) Show that if ψ is any nonprincipal irreducible character of Q, then $\mathrm{Ind}_Q^G(\psi)$ is an irreducible character of G. Show that every irreducible character of G of degree > 1 is equal to $\mathrm{Ind}_Q^G(\psi)$ for some nonprincipal irreducible character ψ of Q. Deduce that every irreducible character of G has degree either 1 or m and the number of irreducible characters of degree m is $(n-1)/m$. [Check that the proof of Proposition 14(3) establishes this more general result with the appropriate changes to the numbers involved.]

8. Use the preceding exercise to explicitly write out the character table of $\langle (1\ 2\ 3\ 4\ 5), (2\ 3\ 5\ 4) \rangle$, which is the normalizer in S_5 of a Sylow 5-subgroup (this group is a Frobenius group of order 20).

9. Let N be a normal subgroup of G, let ψ be a character of N and let $g \in G$. Define ψ^g by $\psi^g(h) = \psi(ghg^{-1})$ for all $h \in N$.
   (a) Prove that ψ^g is a character of N (ψ and ψ^g are called *G-conjugate* characters of N). Prove that ψ^g is irreducible if and only if ψ is irreducible.

(b) Prove that the map $\psi \mapsto \psi^g$ is a right group action of G on the set of characters of N and N is in the kernel of this action.

(c) Prove that if ψ_1 and ψ_2 are G-conjugate characters of N, then $\mathrm{Ind}_N^G(\psi_1) = \mathrm{Ind}_N^G(\psi_2)$. Prove that if ψ_1 and ψ_2 are characters of N that are not G-conjugate then $\mathrm{Ind}_N^G(\psi_1) \neq \mathrm{Ind}_N^G(\psi_2)$. [Use the argument in the proof of Proposition 14(3).]

10. Show that if $G = A_4$ and $N = V_4$ is its Sylow 2-subgroup then any two nonprincipal irreducible characters of N are G-conjugate (cf. the preceding exercise).

11. Let $G = D_{2n}$ be presented by its usual generators and relations. Prove that if ψ is any degree 1 character of $H = \langle r \rangle$ such that $\psi \neq \psi^s$, then $\mathrm{Ind}_H^G(\psi)$ is an irreducible character of D_{2n}. Show that every irreducible character of D_{2n} is the induced character of some degree 1 character of $\langle r \rangle$.

12. Prove both parts of Proposition 15.

13. Prove the following result known as *Frobenius Reciprocity*: let $H \leq G$, let ψ be any character of H and let χ be any character of G. Then

$$(\psi, \chi|_H)_H = (\mathrm{Ind}_H^G(\psi), \chi)_G.$$

[Expand the right hand side using the formula for the induced character $\mathrm{Ind}_H^G(\psi)$.]

14. Assume G were a simple group of order $3^3 \cdot 7 \cdot 13 \cdot 409$ whose Sylow subgroups and their normalizers are described by properties (1) to (5) in this section. Prove that the permutation character of degree 819 obtained from the action of G on the left cosets of the subgroup N_4 decomposes as $\chi_0 + \gamma + \gamma'$, where χ_0 is the principal character of G and γ and γ' are distinct irreducible characters of G of degree 409. [Use Exercise 9 in Section 15.4 to show that this permutation character π has $||\pi||^2 = 3$.]

APPENDIX

Cartesian Products and Zorn's Lemma

Section 1 of the appendix contains the definition of the Cartesian product of an arbitrary collection of sets. In the text we shall primarily be interested in products of finitely many (or occasionally countably many) sets. We indicate how the general definition agrees with the familiar "ordered n-tuple" notion of a Cartesian product in these cases. Section 2 of the appendix contains a discussion of Zorn's Lemma and related topics.

1. CARTESIAN PRODUCTS

A set I is called an *indexing set* or *index set* if the elements of I are used to index some collection of sets. In particular, if A and I are sets, we can form the collection $\{A_i \mid i \in I\}$ by specifying that $A_i = A$ for all $i \in I$. Thus *any* set can be an indexing set; we use this term to emphasize that the elements are used as indices.

Definition

(1) Let I be an indexing set and let $\{A_i \mid i \in I\}$ be a collection of sets. A *choice function* is any function

$$f : I \to \bigcup_{i \in I} A_i$$

such that $f(i) \in A_i$ for all $i \in I$.

(2) Let I be an indexing set and for all $i \in I$ let A_i be a set. The *Cartesian product* of $\{A_i \mid i \in I\}$ is the set of all choice functions from I to $\cup_{i \in I} A_i$ and is denoted by $\prod_{i \in I} A_i$ (where if either I or any of the sets A_i are empty the Cartesian product is the empty set). The elements of this Cartesian product are written as $\prod_{i \in I} a_i$, where this denotes the choice function f such that $f(i) = a_i$ for each $i \in I$.

(3) For each $j \in I$ the set A_j is called the jth *component* of the Cartesian product $\prod_{i \in I} A_i$ and a_j is the jth *coordinate* of the element $\prod_{i \in I} a_i$.

(4) For $j \in I$ the *projection map* of $\prod_{i \in I} A_i$ onto the jth coordinate, A_j, is defined by $\prod_{i \in I} a_i \mapsto a_j$.

Each choice function f in the Cartesian product $\prod_{i \in I} A_i$ may be thought of as a way of "choosing" an element $f(i)$ from each set A_i.

If $I = \{1, 2, \ldots, n\}$ for some $n \in \mathbb{Z}^+$ and if f is a choice function from I to $A_1 \cup \cdots \cup A_n$, where each A_i is nonempty, we can associate to f a unique (ordered) n-tuple:

$$f \to (f(1), f(2), \ldots, f(n)).$$

Note that by definition of a choice function, $f(i) \in A_i$ for all i, so the n-tuple above has an element of A_i in the ith position for each i.

Conversely, given an n-tuple $(a_1, a_1, \ldots, a_n)$, where $a_i \in A_i$ for all $i \in I$, there is a unique choice function, f, from I to $\cup_{i \in I} A_i$ associated to it, namely

$$f(i) = a_i, \qquad \text{for all } i \in I.$$

It is clear that this map from n-tuples to choice functions is the inverse to the map described in the preceding paragraph. Thus *there is a bijection between ordered n-tuples and elements of* $\prod_{i \in I} A_i$. Henceforth when $I = \{1, 2, \ldots, n\}$ we shall write

$$\prod_{i=1}^{n} A_i \quad \text{or} \quad A_1 \times A_2 \times \cdots \times A_n$$

for the Cartesian product and we shall describe the elements as ordered n-tuples.

If $I = \mathbb{Z}^+$, we shall similarly write: $\prod_{i=1}^{\infty} A_i$ or $A_1 \times A_2 \times \cdots$ for the Cartesian product of the A_i's. We shall write the elements as ordered tuples: $(a_1, a_2, \ldots)$, i.e. as infinite sequences whose i^{th} terms are in A_i.

Note that when $I = \{1, 2, \ldots, n\}$ or $I = \mathbb{Z}^+$ we have used the natural ordering on I to arrange the elements of our Cartesian products into n-tuples. Any other ordering of I (or any ordering on a finite or countable index set) gives a different representation of the elements of the same Cartesian product.

Examples

(1) $A \times B = \{(a, b) \mid a \in A, b \in B\}$.

(2) $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ (n factors) is the usual set of n-tuples with real number entries, Euclidean n-space.

(3) Suppose $I = \mathbb{Z}^+$ and A_i is the same set A, for all $i \in I$. The Cartesian product $\prod_{i \in \mathbb{Z}^+} A$ is the set of all (infinite) sequences $a_1, a_2, a_3 \ldots$ of elements of A. In particular, if $A = \mathbb{R}$, then the Cartesian product $\prod_{i \in \mathbb{Z}^+} \mathbb{R}$ is the set of all real sequences.

(4) Suppose I is any indexing set and A_i is the same set A, for all $i \in I$. The Cartesian product $\prod_{i \in I} A$ is just the set of all functions from I to A, where the function $f : I \to A$ corresponds to the element $\prod_{i \in I} f(i)$ in the Cartesian product. This Cartesian product is often (particularly in topology books) denoted by A^I. Note that for each fixed $j \in I$ the projection map onto the jth coordinate sends the function f to $f(j)$, i.e. is evaluation at j.

(5) Let R be a ring and let x be an indeterminate over R. The definition of the ring $R[x]$ of polynomials in x with coefficients from R may be given in terms of Cartesian products rather than in the more intuitive and familiar terms of "formal sums" (in Chapters 7 and 9 we introduced them in the latter form since this is the way we envision and work with them). Let I be the indexing set $\mathbb{Z}^+ \cup \{0\}$ and let $R[x]$ be the subset of the Cartesian product $\prod_{i=0}^{\infty} R$ consisting of elements $(a_0, a_1, a_2, \ldots)$ such

that only finitely many of the a_i's are nonzero. If $(a_0, a_1, a_2, \ldots, a_n, 0, 0, \ldots)$ is such a sequence we represent it by the more familiar "formal sum": $\sum_{i=0}^{n} a_i x^i$. Addition and multiplication of these sequences is defined so that the usual rules for addition and multiplication of polynomials hold.

Proposition 1. Let I be a nonempty countable set and for each $i \in I$ let A_i be a set. The cardinality of the Cartesian product is the product of the cardinalities of the sets A_i, i.e.

$$|\prod_{i \in I} A_i| = \prod_{i \in I} |A_i|.$$

(where if some A_i is an infinite set or if I is infinite and an infinite number of A_i's have cardinality ≥ 2, both sides of this equality are infinity). In particular,

$$|A_1 \times A_2 \times \cdots \times A_n| = |A_1| \times |A_2| \times \cdots \times |A_n|.$$

Proof: In order to count the number of choice functions note that each $i \in I$ may be mapped to any of the $|A_i|$ elements of A_i and for $i \neq j$ the values of choice functions at i and j may be chosen completely independently. Thus the number of choice functions is the product of the cardinalities of the A_i's, as claimed.

For Cartesian products of finitely many sets, $A_1 \times A_2 \times \cdots \times A_n$, one can see this easily from the n-tuple representation: the elements of $A_1 \times A_2 \times \cdots \times A_n$ are n-tuples $(a_1, a_2, \ldots, a_n)$ and each a_i may be chosen as any of the $|A_i|$ elements of A_i. Since these choices are made independently for $i \neq j$, there are $|A_1| \cdot |A_2| \cdots |A_n|$ elements in the Cartesian product.

EXERCISE

1. Let I and J be any two indexing sets and let A be an arbitrary set. For any function $\varphi : J \to I$ define

$$\varphi^* : \prod_{i \in I} A \to \prod_{j \in J} A \qquad \text{by} \qquad \varphi^*(f) = f \circ \varphi \quad \text{for all choice functions } f \in \prod_{i \in I} A.$$

   (a) Let $I = \{1, 2\}$, let $J = \{1, 2, 3\}$ and let $\varphi : J \to I$ be defined by $\varphi(1) = 2$, $\varphi(2) = 2$ and $\varphi(3) = 1$. Describe explicitly how a 3-tuple in $A \times A \times A$ maps to an ordered pair in $A \times A$ under this φ^*.

   (b) Let $I = J = \{1, 2, \ldots, n\}$ and assume φ is a permutation of I. Describe in terms of n-tuples in $A \times A \times \cdots \times A$ the function φ^*.

2. PARTIALLY ORDERED SETS AND ZORN'S LEMMA

We shall have occasion to use Zorn's Lemma as a form of "infinite induction" in a few places in the text where it is desirable to know the existence of some set which is *maximal* with respect to certain specified properties. For example, Zorn's Lemma is used to show that every vector space has a basis. In this situation a basis of a vector space V is a subset of V which is maximal as a set consisting of linearly independent vectors

(the maximality ensures that these vectors span V). For finite dimensional spaces this can be proved by induction; however, for spaces of arbitrary dimension Zorn's Lemma is needed to establish this. By having results which hold in full generality the theory often becomes a little neater in places, although the main results of the text do not require its use.

A specific instance in the text where a maximal object which helps to simplify matters is constructed by Zorn's Lemma is the algebraic closure of a field. An algebraic closure of a field F is an extension of F which is maximal among any collection of algebraic extensions. Such a field contains (up to isomorphism) all elements which are algebraic over F, hence all manipulations involving such algebraic elements can be effected in this one larger field. In any particular situation the use of an algebraic closure can be avoided by adjoining the algebraic elements involved to the base field F, however this becomes tedious (and often obscures matters) in complicated proofs. For the specific fields appearing as examples in this text the use of Zorn's Lemma to construct an algebraic closure can be avoided (for example, the construction of an algebraic closure of any subfield of the complex numbers or of any finite field does not require it).

The first example of the use of Zorn's Lemma appears in the proof of Proposition 11 in Section 7.4.

In order to state Zorn's Lemma we need some terminology.

Definition. A *partial order* on a nonempty set A is a relation $\leq$ on A satisfying

(1) $x \leq x$ for all $x \in A$ (reflexive),

(2) if $x \leq y$ and $y \leq x$ then $x = y$ for all $x, y \in A$ (antisymmetric),

(3) if $x \leq y$ and $y \leq z$ then $x \leq z$ for all $x, y, z \in A$ (transitive).

We shall usually say that A is a partially ordered set under the ordering $\leq$ or that A is partially ordered by $\leq$.

Definition. Let the nonempty set A be partially ordered by $\leq$.

(1) A subset B of A is called a *chain* if for all $x, y \in B$, either $x \leq y$ or $y \leq x$.

(2) An *upper bound* for a subset B of A is an element $u \in A$ such that $b \leq u$, for all $b \in B$.

(3) A *maximal element* of A is an element $m \in A$ such that if $m \leq x$ for any $x \in A$, then $m = x$.

In the literature a chain is also called a *tower* or called a *totally ordered* or *linearly ordered* or *simply ordered* subset.

Some examples below highlight the distinction between upper bounds and maximal elements. Also note that if m is a *maximal* element of A, it is not necessarily the case that $x \leq m$ for all $x \in A$ (i.e. m is not necessarily a *maximum* element).

Examples

(1) Let A be the power set (i.e. set of all subsets) of some set X and $\leq$ be set containment: $\subseteq$. Notice that this is only a *partial* ordering since some subsets of X may not be comparable, e.g. singletons: if $x \neq y$ then $\{x\} \not\subseteq \{y\}$ and $\{y\} \not\subseteq \{x\}$. In this

situation an example of a chain is a collection of subsets of X such as:

$$X_1 \subseteq X_2 \subseteq X_3 \subseteq \cdots$$

Any subset B of A has an upper bound, b, namely,

$$b = \bigcup_{x \in B} x.$$

This partially ordered set A has a (unique) maximal element, X.
In many instances the set A consists of some (but not necessarily all) subsets of a set X (i.e. A is a subset of the power set of X) and with the ordering on A again being inclusion. The existence of upper bounds and maximal elements depends on the nature of A.

(2) Let A be the collection of all *proper* subsets of $\mathbb{Z}^+$ ordered under $\subseteq$. In this situation, chains need not have maximal elements, e.g. the chain

$$\{1\} \subseteq \{1, 2\} \subseteq \{1, 2, 3\} \subseteq \cdots$$

does not have an upper bound. The set A does have maximal elements: for example $\mathbb{Z}^+ - \{n\}$ is a maximal element of A for any $n \in \mathbb{Z}^+$.

(3) Let $A = \mathbb{R}$ under the usual $\leq$ relation. In this example every subset of A is a chain (including A itself). The notion of a subset of A having an upper bound is the same as the usual notion of a subset of $\mathbb{R}$ being bounded above by some real number (so some sets, such as intervals of finite length, have upper bounds and others, such as the set of positive reals, do not). The set A does not have a maximal element.

Zorn's Lemma. If A is a nonempty partially ordered set in which every chain has an upper bound then A has a maximal element.

It is a nontrivial result that *Zorn's Lemma is independent of the usual (Zermelo–Frankel) axioms of set theory*[1] in the sense that if the axioms of set theory are consistent,[2] then so are these axioms together with Zorn's Lemma; and if the axioms of set theory are consistent, then so are these axioms together with the *negation* of Zorn's Lemma. The use of the term "lemma" in Zorn's Lemma is historical.

For the sake of completeness (and to relate Zorn's Lemma to formulations found in other courses) we include two other equivalent formulations of Zorn's Lemma.

The Axiom of Choice. The Cartesian product of any nonempty collection of nonempty sets is nonempty. In other words, if I is any nonempty (indexing) set and A_i is a nonempty set for all $i \in I$, then there exists a choice function from I to $\cup_{i \in I} A_i$.

Definition. Let A be a nonempty set. A *well ordering* on A is a total ordering on A such that every nonempty subset of A has a minimum (or smallest) element, i.e. for each nonempty $B \subseteq A$ there is some $s \in B$ such that $s \leq b$, for all $b \in B$.

[1] See P.J. Cohen, *Proc. Nat. Acad. Sci.*, 50(1963), and 51(1964).
[2] This is not known to be the case!

The Well-Ordering Principle. Every nonempty set A has a well ordering.

Theorem 2. Assuming the usual (Zermelo–Frankel) axioms of set theory, the following are equivalent:

(1) Zorn's Lemma

(2) the Axiom of Choice

(3) the Well-Ordering Principle.

Proof: This follows from elementary set theory. We refer the reader to *Real and Abstract Analysis* by Hewitt and Stromberg, Springer-Verlag, 1965, Section 3 for these equivalences and some others.

EXERCISES

(1) Let A be the collection of all finite subsets of $\mathbb{R}$ ordered by inclusion. Discuss the existence (or nonexistence) of upper bounds, minimal and maximal elements (where minimal elements are defined analogously to maximal elements). Explain why this is not a well ordering.

(2) Let A be the collection of all infinite subsets of $\mathbb{R}$ ordered by inclusion. Discuss the existence (or nonexistence) of upper bounds, minimal and maximal elements. Explain why this is not a well ordering.

(3) Show that the following partial orderings on the given sets are not well orderings:

(a) $\mathbb{R}$ under the usual relation $\leq$.

(b) $\mathbb{R}^+$ under the usual relation $\leq$.

(c) $\mathbb{R}^+ \cup \{0\}$ under the usual relation $\leq$.

(d) $\mathbb{Z}$ under the usual relation $\leq$.

(4) Show that $\mathbb{Z}^+$ is well ordered under the usual relation $\leq$.

Index

A

A.C.C. (See Ascending chain condition)
Abelian extensions of $\mathbb{Q}$, 510 *ff.*, 514
Abelian group, 18, 159 *ff.*, 317
  Fundamental Theorem, 159 *ff.*
  representations of, 584, 585, 597
Action group, 40, 113
  left versus right, 130
Affine algebraic set, 312
Affine n-space, 310, 316
Affords, a representation, 113, 574
Algebra, 572
Algebraically closed, 311, 455, 530
Algebraically conjugate characters, 613 *ff.*
Algebraically independent, 559
Algebraically indistinguishable, 430
Algebraic closure of finite field, 502
Algebraic closures, 455 *ff.*, 502
Algebraic element, 432, 439
Algebraic extension, 432, 439
Algebraic integer, 622 *ff.*
Algorithm for Jordan canonical form, 408
Algorithm for rational canonical form, 393
Alternating function, 361
Alternating group, 108 *ff.*, 151 *ff.*
Angle, trisecting, 446, 447
Annihilated by, 316
Annihilator, of submodule, 371
Annihilators, 248, 321, 360
Artin–Schreier extensions, 503, 550
Ascending chain condition, 288, 369
Associate, 282
Associative, 17
Augmentation ideal, 245, 255, 577
Augmentation map, 245
Augmented matrix, 349
Aut($\mathbb{R}/\mathbb{Q}$), 480
Automorphism, of a field, 468, 471
Automorphism group, 39, 134 *ff.*
  D_8, 137
  Q_8, 137
  S_n, 137
  cyclic group, 136, 137
  elementary abelian group, 137, 323
  of a field, 471
  of a field extension, 471
A_4, characters of, 618
A_5, simplicity of, 129, 151
A_n, 110
  conjugacy in, 132
  simplicity of, 129, 151 *ff.*

B

Base field, 423
Basis, 330, 336
Berlekamp's Factorization Algorithm, 503 *ff.*
Betti number, 159, 374
Bijective, 2
Binary operation, 17
Binary relation, 3
Binomial Theorem, 249
Biquadratic extensions, 442, 496, 503
Biquadratic polynomial, 532
Block, 118
Boolean ring, 231, 248, 257, 270
Building-Up Lemma, 338
Burnside's Basis Theorem, 201
Burnside's Lemma, 612
Burnside's $p^a q^b$ Theorem, 198, 621 *ff.*

C

$\mathbb{C}$, 427, 530
Cancellation laws, 21
Cardano's formulas for roots of a cubic, 543 *ff.*, 545
Cardinality, 1
Cartesian product, 1, 643 *ff.*
Castelnuevo's Theorem, 560
Casus irreducibilis, 547, 551
Cauchy's Theorem, 92, 95, 102
Cayley–Hamilton Theorem, 390
Cayley's Theorem, 121

Center
of group, 49
of p-group, 126, 190
of ring, 231, 572, 592, 596
Central idempotent, 333
Centralize, 94
Centralizer, 48
of representation, 584
Central product, 158
Change of basis, 346
Characteristic:
of a field, 422
of a ring, 249
Characteristic p fields, 461
Characteristic polynomial, 384
Characteristic subgroup, 135, 136, 139
Character of a group, 482, 601
Character table, 615
Chinese Remainder Theorem, 267 *ff.*, 270, 307
for modules, 333
Class equation, 125
Class field theory, 514
Class functions, 601, 605
Classical Greek problems, 446
Closed, 17
Coefficient matrix, 349
Cofactor, 364
Column rank, 345, 359
Comaximal ideals, 268, 280
Commutative, 17, 18
Commutator, 89, 170, 195 *ff.*
Commutator series, 197
Commutator subgroup, 89, 170, 195 *ff.*, 583
Commute, diagram, 101
Companion matrix, 386
Complement, of subgroup, 182, 626
Completely reducible module, 578, 582, 591
Composite extensions, 505 *ff.*
Composites of fields, 440 *ff.*
Composition factors, 103
Composition series, 103
Congruence class, 8
Conjugacy class, 124, 597
Conjugate:
of element, 82, 124
of set, 124
Conjugate field elements, 487
Conjugation, 43, 51
in A_n, 132
in S_n, 126 *ff.*, 133
Constituent, of a module, 578
Constructibility of the regular n-gon, 515 *ff.*
Constructible elements, 444
Construction of cube roots, 447
Construction of the regular 17-gon, 516, 518 *ff.*
Corresponding group actions, 130
Coset, 77
Coset representatives, 77
Cramer's rule, 363
Cubic equations, formulas for roots, 543 *ff.*, 545
Cycle, 28
Cycle decomposition, 28, 116
Cycle decomposition algorithm, 29
Cycle type, 127
Cycle types, of automorphisms, 543, 553
Cyclic extensions, 538, 540
Cyclic group, 53 *ff.*, 160
characters of, 615, 616
representations of, 575, 579, 584, 599
Cyclic module, 327, 373
Cyclic subgroup, 29
Cyclotomic extensions, 464 *ff.*, 510 *ff.*
Cyclotomic fields, 451 *ff.*, 464 *ff.*
Cyclotomic polynomial, 304, 401, 465 *ff.*, 468
Cyclotomy, 512

D

D.C.C. (*See* Descending chain condition)
Decomposable module, 578
Dedekind and Hasse Criterion, 281
Degree:
of a field element, 433
of a field extension, 424, 433
of polynomial, 234, 293
of representation, 571, 597, 603, 624
of symmetric group, 28
Degree 1 representations, 583, 584, 601, 613
Density of primes, 556
Derivative, of a polynomial, 306, 458
Derived series, 197
Descending chain condition, 591
Determinant of a matrix, 247, 360 *ff.*, 366
Diagonalizable matrices, criterion, 405
Dihedral group, 24
as Galois group, 558
Dimension, 335, 338
Diophantine equations, 245, 276

Direct product, 19, 153 *ff.*, 158, 643
of modules, 329, 333, 341
of rings, 267
Direct sum, 158
of modules, 329, 333, 334, 341
of rings, 231
Dirichlet's Theorem on Primes in Arithmetic Progression, 469
Discrete valuation ring, 233, 260, 272
Discriminant, 524
of a cubic, 527
of a polynomial, 524, 533, 536
of pth cyclotomic polynomial, 536
of a quadratic, 526
of a quartic, 529
as resultant, 536
Distributive law, 33, 225
Divisible group, 65, 86, 590
Division Algorithm, 5, 271
Division ring, 226, 252, 256
Double coset, 118
Doubling the cube, impossibility of, 446
Doubly transitive, 118, 613
Dual basis, 357
Dual group, 168
Dual vector space, 357 *ff.*
D_8:
characters of, 609, 614, 616, 633
representations of, 583
D_{10}, characters of, 612, 614
D_{12}, representation of, 632
D_{2n}, 24
commutator subgroup, 172
conjugacy in, 133
characters of, 602, 642
representations of, 576, 579

E

Eigenspace, 384
Eigenvalue, 341, 383
Eigenvector, 341, 383
Eisenstein's Criterion, 303, 304
Elementary abelian group, 137, 156, 317
Elementary Divisor Decomposition Algorithm, 407 *ff.*
Elementary divisors, 162, 375
Elementary Jordan matrix, 404
Elementary row operations, 350, 390
Elementary symmetric functions, 522
Elliptic function field, 566
Embedding of fields, 483, 489
Endomorphism, 324
Endomorphism ring, 324
$\text{End}_R(M)$, 324, 333, 360
Equivalence class, 3
Equivalence relation, 3
Equivalent representations, 578, 602, 604
Euclidean Algorithm, 5, 271, 274, 294
Euclidean Domain, 271, 279, 285, 294
Euclidean Domains, modules over, 381 *ff.*, 401 *ff.*
Euler ϕ-function, 7, 269, 310
Euler's Theorem, 96
Evaluation homomorphism, 244, 358
Even permutation, 108
Exceptional characters, 639
Existence of finite fields, 462
Exponent, 167
Exponentials of matrices, 415
Expressed by radicals, 540
Extension field, 423
Extension theorem for isomorphisms of fields, 431, 454
External direct product, 173

F

Factor group, 76
Factor through, homomorphism, 101
Faithful, 42, 113
Faithful representation, 571, 613
F-algebra, 572
Feit–Thompson Theorem, 104, 198, 636
Fermat primes, 515
Fermat's Little Theorem, 96
Fermat's Theorem on sums of squares, 309
FG-modules, 574, 580, 582
Fiber, 2, 73
Field, 33, 226, 252, 254, 422
Field generated by, 429
Field of fractions, 261 *ff.*, 281
Finite dimensional, 338
Finite extensions, 433, 434, 439
Finite field, existence and uniqueness, 462, 499
Finite fields, 499 *ff.*, 503
Finitely generated, 64, 159, 220, 250, 327, 436, 559, 582
First Orthogonality Relation, 607, 615
Fix, an element, 471
Fixed field, 473, 485
Formal Laurent series, 238
Formally real fields, 442
Formal power series, 238, 257, 289

Fourier analysis, 610
Fractional linear transformations, 480, 560
Frattini's Argument, 195
Frattini subgroup, 201
Free abelian group, 159, 223, 332
Free generators, 220
Free groups, 216 *ff.*
Free module, 316, 330
Free nilpotent group, 223
Free rank, of a module, 374
Frobenius group, 170, 552, 556, 557, 633
  characters of, 633, 641
  as Galois group, 552, 557, 558
Frobenius kernel, 633
Frobenius map, 461, 468, 469, 479, 499, 518
Frobenius Reciprocity, 642
Function field, 559
  elliptic, 566
Fundamental Theorem of Algebra, 457, 530
Fundamental Theorem of Arithmetic, 6, 287
Fundamental Theorem of Finitely Generated Abelian Groups, 159 *ff.*, 199, 379
Fundamental Theorem of Finitely Generated Modules over a P.I.D., 373, 375, 377
Fundamental Theorem of Galois Theory, 481 *ff.*, 487
Fundamental Theorem on Symmetric Functions, 523, 536 *ff.*
$F[x]$-module, 317 *ff.*, 367 *ff.*

G

Galois closure, 508
Galois conjugates, 487
Galois extension, 475, 485, 486, 487
Galois group, 475
  of a biquadratic, 532
  of composite extension, 505, 506, 507, 513
  of a cubic, 527
  of cyclotomic field, 510, 513
  of finite field, 479, 499
  of a quartic, 529, 532
  of $x^3 - 2$, 477, 521
  of $x^6 - 2x^3 - 2$, 537, 558
  of $x^p - 2$, 495
  of $x^p - x - a$, 503
  of $\mathbb{F}_{p^n}$, 479, 499
  of $\mathbb{Q}(\sqrt{2})$, 476
  of $\mathbb{Q}(2^{1/8}, i)$, 490 *ff.*
  of $\mathbb{Q}(2^{1/8}, i)$, over quadratic subfields, 495
  of $\mathbb{Q}(\sqrt{(2+\sqrt{2})(3+\sqrt{3})})$, 498
  of $\mathbb{Q}(\sqrt{2+\sqrt{2}})$, 496
  of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$, 476, 521
  of $\mathbb{Q}(\sqrt{D_1}, \sqrt{D_2})$, 496
  of $\mathbb{Q}(\zeta_5)$, 511
  of $\mathbb{Q}(\zeta_n)$, 510
  of $\mathbb{Q}(\zeta_n + \zeta_n^{-1})$, 517
  of $\mathbb{Q}(\zeta_p)$, 511
  of $\mathbb{Q}(\zeta_{13})$, 512
Galois groups, infinite, 558 *ff.*, 565
Galois groups of polynomials, 476, 520 *ff.*, 558
Galois groups over $\mathbb{Q}$, 553 *ff.*
Gauss' Lemma, 298
Gauss–Jordan elimination, 349 *ff.*
Gaussian integers, 272, 277, 287, 309
Gauss sums, 550
Generalized character, 636
General linear group, 33, 236, 340
General polynomial, 522
Generalized eigenspace, 413
Generators, 26, 53, 60 *ff.*, 220
  of subgroup, 60
  of submodule, 327
Generators for free module, 330
G-invariant, 574
$GL(V)$, 340, 345, 571
$GL_3(\mathbb{F}_2)$, 209 *ff.*, 213
GL_n, 33, 236, 340, 345
Greatest common divisor, 4, 5, 273, 279, 285
Group, 17
  alternating, 108 *ff.*
  cyclic, 53 *ff.*
  dihedral, 24
  divisible, 65, 86
  general linear, 33
  modular, 72, 87, 101
  quasidihedral, 71, 87, 101
  quaternion, 34
  semidihedral (*See* Quasidihedral group)
  special linear, 47
  symmetric, 28
  of units in a ring, 228
Group action, left versus right, 130
Group ring, 236 *ff.*, 571
Groups:
  of medium order, 203 *ff.*

of order $3^3 \cdot 7 \cdot 13 \cdot 409$, 214, 216, 636 *ff.*
of order 12, 145, 184
of order 30, 145, 183
of order 60, 146
of order pq, 136, 144, 183
of order p^2q, 146
of order p^3, 175, 185
simple of order 168, 209 *ff.*, 216
table of small order, 169
G-stable, 574

H

Hall's Theorem, 198, 626
Hall subgroup, 102, 202, 626
Hamilton Quaternions, 226, 230, 231, 237, 257
Harmonic analysis, 610
Heptadecagon, construction of, 518 *ff.*
Hilbert's Basis Theorem, 289
Hilbert's Nullstellensatz, 312
Hilbert's Specialization Theorem, 562
Hilbert's Theorem, 90, 497
additive form, 497
Hölder program, 104, 585
Holomorph, 180, 186
Homogeneous, 536
Homomorphism:
group, 35
of modules, 322 *ff.*
ring, 239
$\mathrm{Hom}_F(V, W)$, 343
$\mathrm{Hom}_R(M, M)$, 322
Hypernilpotent group, 193

I

Ideal, 242
generated by set, 250
Idempotent, 269, 333, 349, 592, 604, 606
Identity:
group, 18
ring, 225
Image of a map, 2
Image of linear transformation, computing, 354
Indecomposable module, 578, 580
Index, of subgroup, 90
Induced character, 628 *ff.*, 630, 635
Induced homomorphism, 101
Induced module, 628 *ff.*, 630
Induced representation, 630 *ff.*
Infinite Galois groups, 558 *ff.*
Injective, 2
Injective module, 585 *ff.*, 590
Inner automorphism, 135
Inner product, of characters, 605
Inseparable degree:
of field extension, 563
of a polynomial, 463
Inseparable extension, 463, 558 *ff.*
Inseparable polynomial, 458
Insolvability of the quintic, 542, 552
Integral domain, 229, 235, 255, 256, 264, 265, 282
Internal direct product, 173
Intertwine, 578
Invariant Factor Decomposition Algorithm, 391, 393
Invariant factors, 159, 374
for a matrix, 388, 414
Invariant subspace, 319
Inverse, 18
Inverses of matrices, computing, 352, 365
Irreducibility, of cyclotomic polynomial, 304, 466
Irreducibility criteria, 301 *ff.*
Irreducible character, 601, 605, 607
Irreducible element, 282 *ff.*
Irreducible module, 333, 578, 583
Irreducible polynomial, 301
Irreducible polynomials of degree n over $\mathbb{F}_p$, 500, 502
Isomorphism:
group, 35
of modules, 322
of rings, 239
Isomorphism Theorems:
for groups, 96 *ff.*
for modules, 326
for rings, 234, 246
Isotypic component, 605, 612

J

Jacobson radical, 258
Join, 65
Jordan block, 404
Jordan canonical form, 402 *ff.*, 404, 405, 407 *ff.*
Jordan–Hölder Theorem, 103

K

Kernel:
of action, 42, 50, 133

Kernel (*cont.*)
of homomorphism, 39, 75
of module homomorphism, 322
ring homomorphism, 239
Kernel of linear transformation, computing, 354
Klein 4-group, 68
Kronecker–Weber Theorem, 514, 536
Krull topology, 565
k-stage Euclidean Domains, 289
Kummer extensions, 540
Kummer generators for cyclic extensions, 550
Kummer theory, 540

L

Lagrange resolvent, 539
Lagrange's Theorem, 44, 89
Lattice:
of subgroups, 65 *ff.*
of subgroups of A_4, 111
of subgroups of D_8, 69, 100
of subgroups of D_{16}, 70
of subgroups of QD_{16}, 72
of subgroups of Q_8, 69, 99
of subgroups of S_3, 68
of subgroups of $\mathbb{Z}/2\mathbb{Z}$, 66
of subgroups of $\mathbb{Z}/4\mathbb{Z}$, 66
of subgroups of $\mathbb{Z}/6\mathbb{Z}$, 67
of subgroups of $\mathbb{Z}/8\mathbb{Z}$, 67
of subgroups of $\mathbb{Z}/12\mathbb{Z}$, 68
of subgroups of $\mathbb{Z}/p^n\mathbb{Z}$, 67
of subgroups of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, 68
of subgroups of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/4\mathbb{Z}$, 72
of subgroups of $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$, 72
of subgroups of Klein 4-group, 68
of subgroups of modular group, 72
of subgroups of quasidihedral group, 493
of subfields, 488
of subfields of $\mathbb{Q}(2^{1/8}, i)$, 495
of subfields of $\mathbb{Q}(\zeta_{13})$, 512
of subfields of $\mathbb{Q}(\sqrt[3]{2}, \rho)$, 481
of subfields of $\mathbb{Q}(\sqrt{2}, \sqrt{3})$, 481
Lattice for quotient group, 99 *ff.*
Laurent series, 238
Least common multiple, 4, 278, 288
Least residue, 9
Left regular representation, 121
Length, of cycle, 29
Lexicographic ordering, 536
Linear equations, solving, 351
Linear functionals, 357
Linear independence of characters, 482
Linearly independent, 336, 353, 370
Linear representation, 571
Linear transformation, 323, 335
Localization, 266, 267
Local ring, 260
Lower central series, 196
Lüroth's Theorem, 560

M

Maschke's Theorem, 580, 586
Matrix, of a linear transformation, 341 *ff.*
Matrix representation, 571
Matrix ring, 235
Maximal ideal, 253 *ff.*, 256, 258, 280
Maximal real subfield, of cyclotomic field, 517
Maximal subgroup, 64, 118, 190
Maximal subgroups, of solvable groups, 202
Minimal normal subgroup, 202
Minimal polynomial, 385
of a field element, 433
Minimum condition, 591
Minor, 364
Modular arithmetic, 9
Modular group, 72, 87, 101, 188
Modular representations, 577
Module, 315
over group ring, 574
over $F[x]$, 317 *ff.*, 367
over $\mathbb{Z}$, 317, 367
Monic, 234
Monomial, 234, 293
Monster simple group, 600, 615
Multilinear, 361
Multiple root, of polynomial, 306, 307
Multiple roots, 458
Multiplicativity of extension degrees, 435
Multiplicity, of root, 458
Möbius inversion formula, 502

N

Natural projection, 83, 243, 325
n-cycle, 29, 117
holomorph of, 188

Newton's formulas, 533
Nilpotence class, 192
Nilpotent element, 231, 249
Nilpotent group, 192
Nilpotent ideal, 250
Nilpotent matrix, 413
Nilradical, 249, 258, 267
Noetherian modules, 369, 380
Noetherian ring, 288
Nonsingular, 340, 343
Norm, 232, 271, 496
  of a character, 608
  of element in a field, 496
Normal extension, 449, 564, 565
Normalize, 94
Normalizer, 49
Normal subgroup, 82
Null space, 340

O

Odd permutation, 108
Orbit, 44, 116, 612
Order:
  of an element, 21
  of permutation, 31
  of a set, 1
Ordered basis, 336
Outer automorphism, 138

P

Partition, 3, 127, 162
Pell's equation, 288
Perfect fields, 462, 463
Perfect group, 176
Periods, in cyclotomic fields, 512, 518
Permutation, 3, 28
Permutation character, 602, 608, 609, 633
Permutation groups, 117
Permutation matrix, 157
Permutation representation, 41, 113, 115
p-extensions, 510
p-group, 140, 190 *ff.*
p-groups, characters of group of order p^3, 621
p-groups, representations of, 585, 599
P.I.D. (*See* Principal Ideal Domain)
Pivotal element, 350
Polynomial ring, 234 *ff.*, 291 *ff.*
Polynomials with S_n, as Galois group, 555
Power, of ideal, 250
Power series, 238, 257, 289
Power series of matrices, 414
Preimage, 2
Presentation, 27, 220
Primary component, 144, 333, 334, 376
Primary Decomposition Theorem, 162, 376
Primary ideal, 260, 294
Prime element, 282 *ff.*
Prime factorization, 283
Prime ideal, 254, 267, 280
Prime spectrum, 260
Prime subfield, 266, 423
Primitive, 118
Primitive element, 429
Primitive Element Theorem, 509
Primitive roots of unity, 452, 465
Principal character, 601
Principal ideal, 250, 273
Principal Ideal Domain, 279 *ff.*, 285, 294, 328, 333, 334, 367 *ff.*, 586, 589
Principal Ideal Domain, that is not Euclidean, 280
Product, of ideals, 250, 256, 268
Projection, 83, 154, 325, 349, 581
Projective module, 585 *ff.*, 590
Public key code, 278
Purely inseparable, 563
Purely transcendental, 560

Q

QD_{2n}, 71, 87, 101, 188, 493
Quadratic extensions, 434
Quadratic subfields of cyclic quartic fields, criterion, 552
Quartic equations, formulas for roots, 548 *ff.*
Quasidihedral group, 71, 87, 101, 188, 493
  as Galois group, 493
Quaternion group, 34, 69, 125
  as Galois group, 498
  generalized, 180
Quaternion ring, 226, 230, 231, 237, 257
Quintic, insolvability, 542, 552
Quotient field, 261 *ff.*
Quotient group, 73, 76
Quotient module, 325
Quotient ring, 239 *ff.*

Quotient space, 335
Q_8, 34, 69, 125
characters of, 602, 612, 614, 617
representations of, 576, 584

R

Radical, 258, 267
of ideal, 289
Radical extensions, 538 *ff.*
Rank:
of a free module, 316, 330
of a group, 166, 220
of linear transformation, 340
of a module, 371, 379, 380
Rational canonical form, 383 *ff.*, 386, 387, 388, 393
Rational functions, 265, 428, 443
Rational valued characters, criterion for, 614
Reduced row echelon form, 350
Reducible element, 282
Reducible module, 578, 580
Reduction homomorphism, 245
Regular character, 602, 633
Regular representation, 121, 575, 579, 585, 599
Relations, 26, 220
Relations matrix, 381
Relative degree, of a field extension, 424
Relatively prime, 4
Replacement Theorem, for vector spaces, 337
Representation:
linear, 571
matrix, 571
permutation, 41
Representatives, 9, 75, 77
Residue class, 8
Resolvent cubic, 529
Resolvent polynomials, 529, 537, 548, 552, 556, 557
Restricted direct product, 158
Resultant, 534 *ff.*
Right group actions, 130
Ring, 225
Ring of algebraic integers, 628
Ring of fractions, 261 *ff.*, 281
R-module, 315
Root, of polynomial, 301
Roots of unity, 23, 451, 464
Row equivalent, 350
Row rank, 345, 359
Row reduced, 350
Ruler and compass constructions, 447

S

Scalar, 335
Scalar matrix, 236
Scalar transformations, 324
Schur's Lemma, 333, 584, 585, 593
Second Orthogonality Relation, 608, 615
Semidihedral group (*See* Quasidihedral group)
Semidirect product, 176 *ff.*
Semisimple, 591
Separability, criterion for, 459 *ff.*
Separable degree:
of a polynomial, 463
of field extension, 563
Separable extension, 463
Separable polynomial, 458 *ff.*
Separating transcendence base, 564, 566
Sign, of permutation, 108
Similar linear transformations, 387
Similar matrices, 346
criterion, 389
Similar representations, 578
Simple extensions, 429, 505 *ff.*, 508, 509, 512
Simple group, 91, 103
Simple groups, classification of, 104, 214
Simple module, 578, 591, 593
Skew field, 226
SL_n, 47, 104
Smith normal form, 391, 414
Solution of cubic equations, 543 *ff.*
Solution of quartic equations, 548 *ff.*
Solvability of quintic, criterion, 552
Solvable extensions, 538 *ff.*, 552
Solvable group, 105, 196 *ff.*, 621, 626, 628
Solved by radicals, 538, 540, 552
Solving linear equations, 351
Span, 328, 335, 353
Special linear group, 47
Specialization, 562
Splitting field, 425, 448 *ff.*, 455, 486
of $x^2 + 1$, 427
of $x^2 + x + 1$ over $\mathbb{F}_2$, 428
of $x^2 - t$ over $k(t)$, 428
of $x^2 - 2$, 449
of $(x^2 - 2)(x^2 - 3)$, 449
of $x^3 - 2$, 449

of $x^4 + 4$, 449
of $x^n - 1$, 451
of $x^p - 2$, 453
of $x^4 + px + p$, 532
of $x^4 - 2x^2 - 2$, 496
of $x^4 - px^2 + q$, 532
of $x^6 - 2x^3 - 2$, 537, 558
of $x^8 - 2$, 490
of $x_p - x - a$ over $\mathbb{F}_p$, 503
Sporadic simple group, 104
Square root of a matrix, 413
Squaring the circle, impossibility of, 446
Stabilizer, 43, 50, 113
Stable subspace, 319
Straightedge and compass constructions, 443 *ff.*
Subgroup, 23, 45
generators of, 61
maximal, 64
normal, 82
Subgroup criterion, 46
Submodule, 315
Submodule Criterion, 320
Subring, 230
Subspace, 335
Sum:
of ideals, 250
of submodules, 325
Surjective, 2
Sylow's Theorem, 93, 140 *ff.*
Sylow subgroup, 102, 140, 162
Symmetric functions, 361, 523 *ff.*
Symmetric group, 28
S_3, characters of, 608, 616, 633
S_3, representations of, 598
S_4, characters of, 609, 618
S_5, characters of, 619
S_n, 28
as Galois group, 562
conjugacy in 126, *ff.*, 133
characters of, 613
representations of, 575, 579, 583, 584, 602
S_p:
normalizer of Sylow p-subgroup, 133, 206
Sylow p-subgroups of, 206
S_{p^2}, Sylow p-subgroups of, 189

T

Table, group, 22
Tchebotarov Density Theorem, 556
Thompson's Theorem, 198
Thompson subgroup, 140
Torsion element, 321
Torsion free, 371
Torsion module, 332
Torsion subgroup, 47
Torsion submodule, 321, 370
Trace:
of element in a field, 497
of matrix, 247, 356, 533
Trace ideal, of group ring, 577
Trace submodule, 577
Trascendence base, 559
Trascendence degree, 559
Trascendental, 432, 559
Transcendental extensions, 558 *ff.*
Transition matrix, 346
Transitive, 116
Transitive subgroup of S_n, 521, 554
Transitive subgroups of S_5, 556
Transpose, 359
Transposition, 107
Trisecting an angle, impossibility of, 446
Trisecting an angle using ruler, 447
Trivial representation, 575
Twisted polynomial ring, 297

U

U.F.D. (*See* Unique Factorization Domain)
Unique Factorization Domain, 282 *ff.*, 294, 299, 300
Uniqueness of finite fields, 462
Uniqueness of splitting field, 455
Units, 228, 235, 237, 252
Units in $\mathbb{Z}/n\mathbb{Z}$, 269, 308
Universal property of free groups, 217
Universal property of free modules, 331
Universal side divisor, 281
Upper central series, 192
Upper triangular matrices, 48

V

Valuation ring, 232
Vandermonde determinant, 534
Vector space, 315, 335 *ff.*
Viergruppe, 68
Virtual character, 638
V_4, 68

W

Wedderburn components, 591
Wedderburn decomposition, 591
Wedderburn's Theorem on Finite Division Rings, 469
Wedderburn's Theorem on Semisimple Rings, 590
Weight, of monomial, 536
Well defined, 2
Well Ordering of $\mathbb{Z}$, 4
Wilson's Theorem, 464
Word, 217, 218
Wreath product, 189

Z

Zariski closed set, 260, 261
Zariski topology, 260, 312
Zero divisors, 228, 233
$\mathbb{Z}$-modules, 317, 323, 367
$\mathbb{Z}[\sqrt{D}]$, 232, 277 *ff.*, 287, 288
Z_n, 56
Zorn's Lemma, 253, 645 *ff.*