

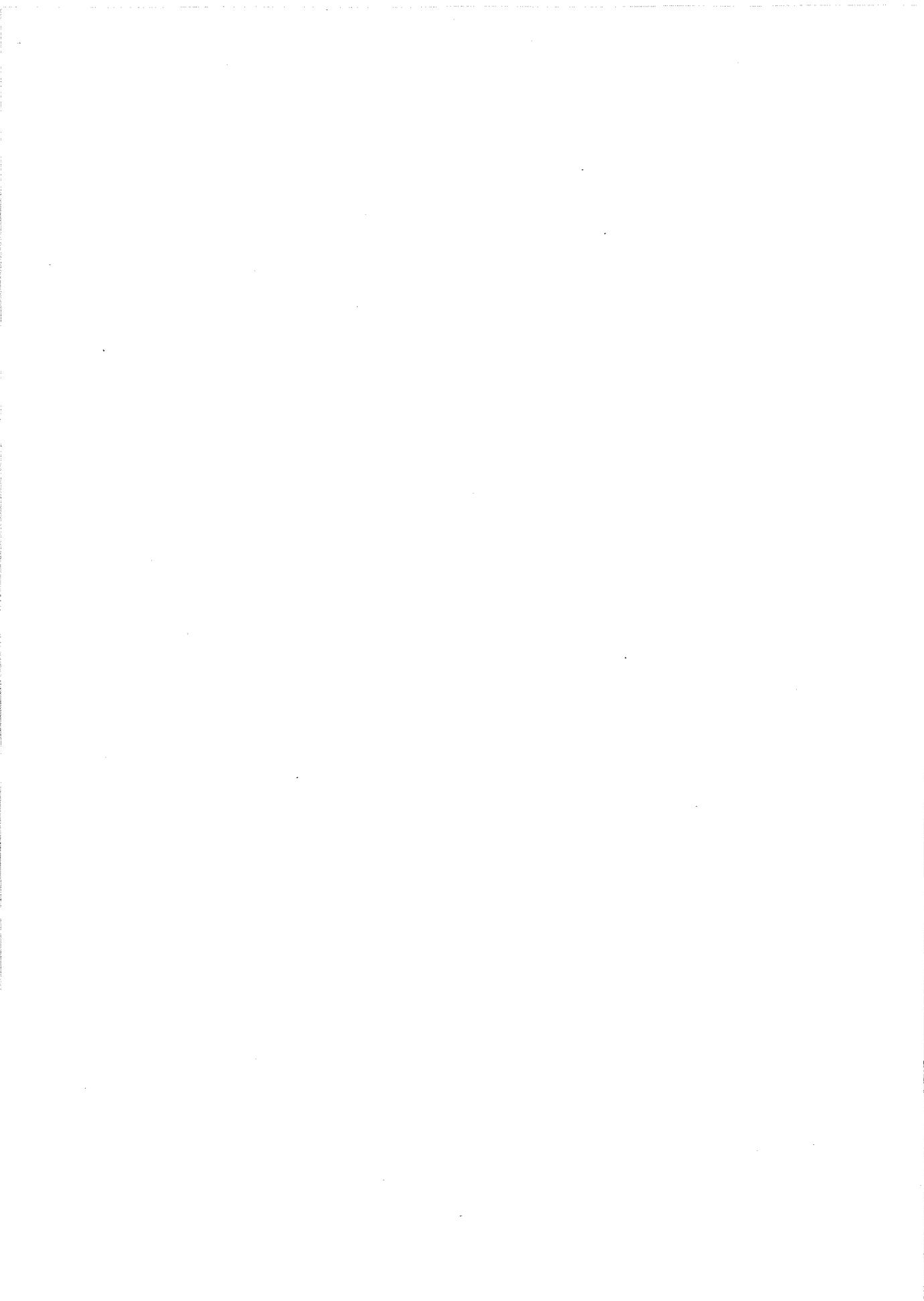
Introduction to ABSTRACT ALGEBRA

Fourth Edition

W. Keith Nicholson

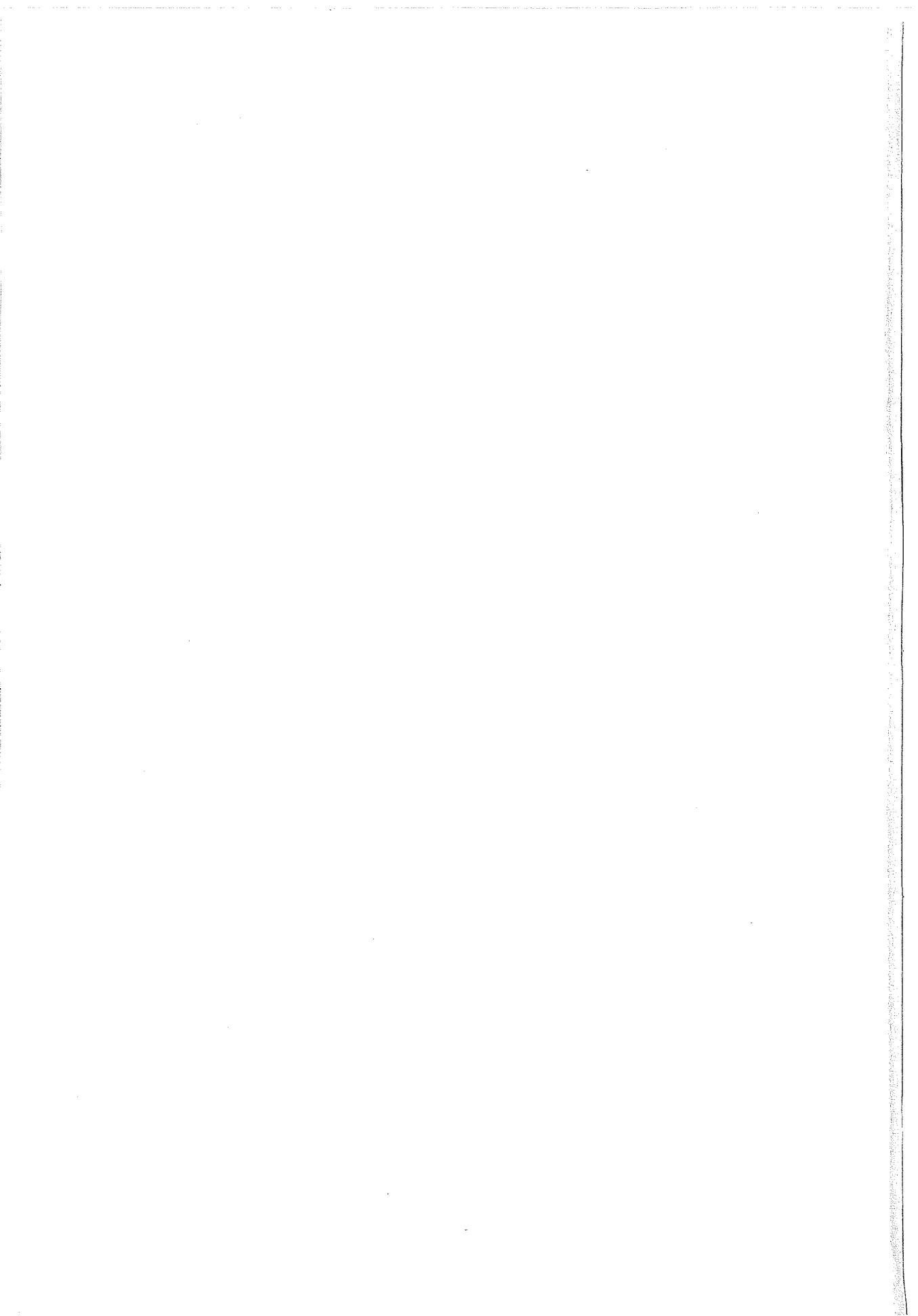


WILEY



*Introduction to
Abstract Algebra*





Introduction to Abstract Algebra

Fourth Edition

W. Keith Nicholson

University of Calgary
Calgary, Alberta, Canada



A JOHN WILEY & SONS, INC., PUBLICATION

Copyright 2012 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department with the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Nicholson, W. Keith.

Introduction to abstract algebra / W. Keith Nicholson. – 4th ed.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-118-13535-8 (cloth)

1. Algebra, Abstract. I. Title.

QA162.N53 2012

512'.02–dc23

2011031416

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Contents

PREFACE	ix
ACKNOWLEDGMENTS	xvii
NOTATION USED IN THE TEXT	xix
A SKETCH OF THE HISTORY OF ALGEBRA TO 1929	xxiii
0 Preliminaries	1
0.1 Proofs / 1	
0.2 Sets / 5	
0.3 Mappings / 9	
0.4 Equivalences / 17	
1 Integers and Permutations	23
1.1 Induction / 24	
1.2 Divisors and Prime Factorization / 32	
1.3 Integers Modulo n / 42	
1.4 Permutations / 53	
1.5 An Application to Cryptography / 67	
2 Groups	69
2.1 Binary Operations / 70	
2.2 Groups / 76	
2.3 Subgroups / 86	
2.4 Cyclic Groups and the Order of an Element / 90	

2.5 Homomorphisms and Isomorphisms / 99	
2.6 Cosets and Lagrange's Theorem / 108	
2.7 Groups of Motions and Symmetries / 117	
2.8 Normal Subgroups / 122	
2.9 Factor Groups / 131	
2.10 The Isomorphism Theorem / 137	
2.11 An Application to Binary Linear Codes / 143	
3 Rings	159
3.1 Examples and Basic Properties / 160	
3.2 Integral Domains and Fields / 171	
3.3 Ideals and Factor Rings / 180	
3.4 Homomorphisms / 189	
3.5 Ordered Integral Domains / 199	
4 Polynomials	202
4.1 Polynomials / 203	
4.2 Factorization of Polynomials Over a Field / 214	
4.3 Factor Rings of Polynomials Over a Field / 227	
4.4 Partial Fractions / 236	
4.5 Symmetric Polynomials / 239	
4.6 Formal Construction of Polynomials / 248	
5 Factorization in Integral Domains	251
5.1 Irreducibles and Unique Factorization / 252	
5.2 Principal Ideal Domains / 264	
6 Fields	274
6.1 Vector Spaces / 275	
6.2 Algebraic Extensions / 283	
6.3 Splitting Fields / 291	
6.4 Finite Fields / 298	
6.5 Geometric Constructions / 304	
6.6 The Fundamental Theorem of Algebra / 308	
6.7 An Application to Cyclic and BCH Codes / 310	
7 Modules over Principal Ideal Domains	324
7.1 Modules / 324	
7.2 Modules Over a PID / 335	

8	<i>p</i>-Groups and the Sylow Theorems	349
8.1	Products and Factors / 350	
8.2	Cauchy's Theorem / 357	
8.3	Group Actions / 364	
8.4	The Sylow Theorems / 371	
8.5	Semidirect Products / 379	
8.6	An Application to Combinatorics / 382	
9	Series of Subgroups	388
9.1	The Jordan–Hölder Theorem / 389	
9.2	Solvable Groups / 395	
9.3	Nilpotent Groups / 401	
10	Galois Theory	412
10.1	Galois Groups and Separability / 413	
10.2	The Main Theorem of Galois Theory / 422	
10.3	Insolvability of Polynomials / 434	
10.4	Cyclotomic Polynomials and Wedderburn's Theorem / 442	
11	Finiteness Conditions for Rings and Modules	447
11.1	Wedderburn's Theorem / 448	
11.2	The Wedderburn–Artin Theorem / 457	
Appendices		471
Appendix A	Complex Numbers / 471	
Appendix B	Matrix Algebra / 478	
Appendix C	Zorn's Lemma / 486	
Appendix D	Proof of the Recursion Theorem / 490	
BIBLIOGRAPHY		492
SELECTED ANSWERS		495
INDEX		523

Preface

This book is a self-contained introduction to the basic structures of abstract algebra: groups, rings, and fields. It is designed to be used in a two-semester course for undergraduates or a one-semester course for seniors or graduates. The table of contents is flexible (see the chapter summaries that follow), so the book is suitable for a traditional course at various levels or for a more application-oriented treatment. The book is written to be read by students with little outside help and so can be used for self-study. In addition, it contains several optional sections on special topics and applications.

Because many students will not have had much experience with abstract thinking, a number of important concrete examples (number theory, integers modulo n , permutations) are introduced at the beginning and referred to throughout the book. These examples are chosen for their importance and intrinsic interest and also because the student can do actual computations almost immediately even though the examples are, in the student's view, quite abstract. Thus, they provide a bridge to the abstract theory and serve as prototype examples of the abstract structures themselves. As an illustration, the student will encounter composition and inverses of permutations before having to fit these notions into the general framework of group theory.

The axiomatic development of these structures is also emphasized. Algebra provides one of the best illustrations of the power of abstraction to strip concrete examples of nonessential aspects and so to reveal similarities between ostensibly different objects and to suggest that a theorem about one structure may have an analogue for a different structure. Achieving this sort of facility with abstraction is one of the goals of the book. This goes hand in hand with another goal: to teach the student how to do proofs. The proofs of most theorems are at least as important for the techniques as for the theorems themselves. Hence, whenever possible, techniques are introduced in examples before giving them in the general case as a proof. This partly explains the large number of examples (over 450) in the book.

Of course, a generous supply of exercises is essential if this subject is to have a lasting impact on students, and the book contains more than 1450 exercises (many with separate parts). For the most part, computational exercises appear first, and the exercises are given in ascending order of difficulty. Hints are given for the less straightforward problems, and answers are provided to odd numbered (parts of) computational exercises and to selected theoretical exercises. (A student solution manual is available.) While exercises are vital to understanding this subject, they are not used to develop results needed later in the text.

An increasing number of students of abstract algebra come from outside mathematics and, for many of them, the lure of pure abstraction is not as strong as for mathematicians. Therefore, applications of the theory are included that make the subject more meaningful and lively for these students (and for the mathematicians!). These include cryptography, linear codes, cyclic and BCH codes, and combinatorics, as well as “theoretical” applications within mathematics, such as the impossibility of the classical geometric constructions. Moreover, the inclusion of short historical notes and biographies should help the reader put the subject into perspective. In the same spirit, some classical “gems” appear in optional sections (one example is the elegant proof of the fundamental theorem of algebra in Section 6.6, using the structure theorem for symmetric polynomials). In addition, the modern flavor of the subject is conveyed by mentioning some unsolved problems and recent achievements, and by occasionally stating more advanced theorems that extend beyond the results in the book.

Apart from that the material is quite standard. The aim is to reveal the basic facts about groups, rings, and fields and give the student the working tools for applications and further study. The level of exposition rises slowly throughout the book and no prior knowledge of abstract algebra is required. Even linear algebra is not needed. Except for a few well-marked instances, the aspects of linear algebra that are needed are developed in the text. Calculus is completely unnecessary. Some preliminary topics that are needed are covered in Chapter 0, with appendices on complex numbers and matrix algebra (over a commutative ring).

Although the chapters are necessarily arranged in a linear order, this is by no means true of the contents, and the student (as well as the instructor) should keep the chapter dependency diagram in mind. A glance at that diagram shows that while Chapters 1–4 are the core of the book, there is enough flexibility in the remaining chapters to accommodate instructors who want to create a wide variety of courses. The jump from Chapter 6 to Chapter 10 deserves mention. The student has a choice at the end of Chapter 6: either change the subject and return to group theory or continue with fields in Chapter 10 (solvable groups are adequately reviewed in Section 10.3, so Chapter 9 is not necessary). The chapter summaries that follow, and the chapter dependency diagram, can assist in the preparation of a course syllabus.

Our introductory course at Calgary of 36 lectures touches Sections 0.3 and 0.4 lightly and then covers Chapters 1–4 except for Sections 1.5, 2.11, 3.5, and 4.4–4.6. The sequel course (also 36 lectures) covers Chapters 5, 6, 10, 7, 8, and 9, omitting Sections 6.6, 6.7, 8.5, 8.6, and 10.4 and Chapter 11.

FEATURES

This book offers the following significant features:

- Self-contained treatment, so the book is suitable for self-study.
- Preliminary material for self-study or review available in Chapter 0 and in Appendices A and B.
- Elementary number theory, integers modulo n , and permutations done first as a bridge to abstraction.
- Over 450 worked examples to guide the student.
- Over 1450 exercises (many with parts), graded in difficulty, with selected answers.
- Gradual increase in level throughout the text.
- Applications to number theory, combinatorics, geometry, cryptography, coding, and equations.
- Flexibility in syllabus construction and choice of optional topics (see chapter dependency diagram).
- Historical notes and biographies.
- Several special topics (for example, symmetric polynomials, nilpotent groups, and modules).
- Solution manual containing answers or solutions to all exercises.
- Student solution manual available with solutions to all odd numbered (parts of) exercises.

CHANGES IN THE THIRD EDITION (2007)

The important concept of a module was introduced and used in Chapters 7 and 11.

- Chapter 7 on finitely generated abelian groups was completely rewritten, modules were introduced, direct sums were studied, and the rank of a free module was defined (for commutative rings). Then the structure of finitely generated modules over a PID was determined.
- Chapter 11 was upgraded from finite dimensional algebras to rings with the descending chain condition. Wedderburn's characterization of simple artinian rings and the Wedderburn–Artin theorem on semisimple rings were proved.
- A new section on semidirect products of groups was added.
- Appendices on Zorn's lemma and the recursion theorem were added.
- More solutions to theoretical exercises were included in the Selected Answers section.

CHANGES IN THE FOURTH EDITION

The changes in the Third Edition primarily involved new concepts (modules, semi-direct products, etc). However, the changes in the Fourth Edition are more “microscopic” in nature, having more to do with clarity of exposition and making

the “flow” of arguments more natural and inevitable. Of course, minor editorial changes are made through the book to correct typographical errors, improve the exposition, and in some cases remove unnecessary material. Here are some more specific changes.

- Because of the increasing importance of modules in the undergraduate curriculum, the new material on modules over a PID (Chapter 7) and the Wedderburn theorems (Chapter 11) introduced in the Third Edition was thoroughly reviewed for clarity of exposition.
- More generally, in an effort to make the book more accessible to students, the writing was carefully edited to ensure readability and clarity, the goal being to make arguments flow naturally and, as much as possible, effortlessly. Of course, this is in accord with the goal of making the book more suitable for self-study.
- Appendix B is expanded to an exposition of matrix algebra over a commutative ring.
- Two notational changes are introduced. First, the symbol $o(g)$ replaces $|g|$ for the order of an element g in a group, reducing confusion with the cardinality $|X|$ of a set. Second, polynomials $f(x)$ are written simply as f .
- In Chapter 2, proofs of two early examples of “structure theorems” are given to motivate the subject: A group of order $2p$ (p a prime) is cyclic or dihedral, and an abelian group of order p^2 is C_{p^2} or $C_p \times C_p$.
- More emphasis is placed on characteristic subgroups and on the product HK of subgroups H and K .
- Wilson’s theorem is included in §1.3 with later applications to number theory and fields.
- In Chapter 5, it is shown that an integral domain is a UFD if and only if it has the ACC on principal ideals and either (a) every irreducible is prime, or (b) any two nonzero elements have a greatest common divisor. This shortens the original proof (with (a) only) at the expense of a lemma of independent interest.
- In Chapter 6, a simpler proof is given that any finite multiplicative subgroup of a field is cyclic.
- The first section of Chapter 8 has been completely rewritten with several results added.
- In Chapter 9, several new results on nilpotent groups have been included. In particular, the Fitting subgroup of any finite group G is introduced, several properties are deduced, and its relationship to the Frattini subgroup is explained.
- In Chapter 10, many arguments are rewritten and clarified, in particular the lemma explaining the basic Galois connection between the subgroups of the Galois group of a field extension and the intermediate fields of the extension.
- In Chapter 11, a new elementary proof is given that $R = L^n$, where L is a simple left ideal of the simple ring R . This directly leads to Wedderburn’s theorem, and the proof does not involve the theory of semisimple modules.
- A student solution manual is now available giving detailed solutions to all odd numbered (parts of) exercises.

CHAPTER SUMMARIES

Chapter 0. Preliminaries. This chapter should be viewed as a primer on mathematics because it consists of materials essential to any mathematics major. The treatment is self-contained. I personally ask students to read Sections 0.1 and 0.2, and I touch briefly on the highlights of Sections 0.3 and 0.4. (Our students have had complex numbers and one semester of linear algebra, so a review of Appendices A and B is left to them.)

Chapter 1. Integers and Permutations. This chapter covers the fundamental properties of the integers and the two prototype examples of rings and groups: the integers modulo n and the permutation group S_n . These are presented naively and allow the students to begin doing ring and group calculations in a concrete setting.

Chapter 2. Groups. Here, the basic facts of group theory are developed, including cyclic groups, Lagrange's theorem, normal subgroups, factor groups, homomorphisms, and the isomorphism theorem. The simplicity of the alternating groups A_n is established for $n \geq 5$. An optional application to binary linear codes is included.

Chapter 3. Rings. The basic properties of rings are developed: integral domains, characteristic, rings of quotients, ideals, factor rings, homomorphisms and the isomorphism theorem. Simple rings are studied, and it is shown that the ring of $n \times n$ matrices over a division ring is simple.

Chapter 4. Polynomials. After the usual elementary facts are developed, irreducible polynomials are discussed and the unique factorization of polynomials over a field is proved. The factor rings of polynomials over a field are described in detail, and some finite fields are constructed. In an optional section, symmetric polynomials are discussed and the fundamental structure theorem is proved.

Chapter 5. Factorization in Integral Domains. Unique factorization domains (UFDs) are characterized in terms of irreducibles, primes, and greatest common divisors. The fact that being a UFD is inherited by polynomial rings is derived. Principal ideal domains and euclidean domains are discussed. This chapter is self-contained, and the material presented is not required elsewhere.

Chapter 6. Fields. After a minimal amount of vector space theory is developed, splitting fields are constructed and used to completely describe finite fields. This topic is a direct continuation of Section 4.3. In optional sections, the classical results on geometric constructions are derived, the fundamental theorem of algebra is proved, and the theory of cyclic and BCH codes is developed.

Chapter 7. Modules over Principal Ideal Domains. Motivated by vector spaces (Section 6.1) and abelian groups, the idea of a module over a ring is introduced. Free modules are discussed and the uniqueness of the rank is proved for IBN rings. With abelian groups as the motivating example, the structure of finitely generated modules over a principal ideal domain is determined, yielding the fundamental theorem for finitely generated abelian groups.

Chapter 8. p -Groups and the Sylow Theorems. This chapter is a direct continuation of Section 2.10. After some preliminaries (including the correspondence theorem), the class equation is developed and used to prove Cauchy's theorem and to derive the basic properties of p -groups. Then group actions are introduced, motivated by the class equation and an extended Cayley theorem, and used to prove

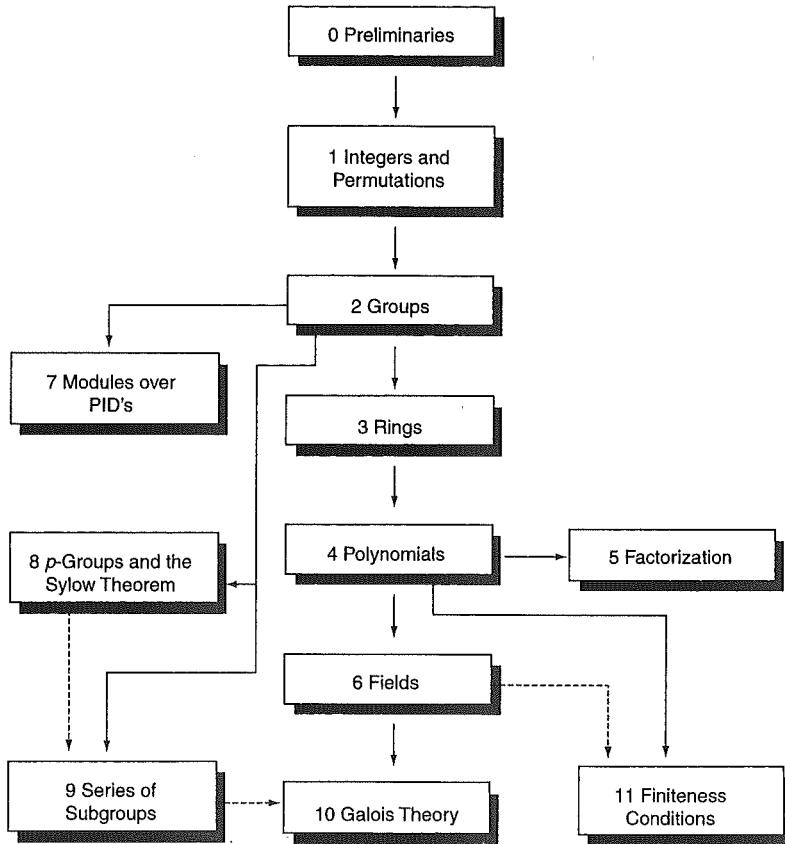
the Sylow theorems. Semidirect products are presented. An optional application to combinatorics is also included.

Chapter 9. Series of Subgroups. The chapter begins with composition series and the Jordan–Hölder theorem. Then solvable series are introduced, including the derived series, and the basic properties of solvable groups are developed. Sections 9.1 and 9.2 depend only on the second and third isomorphism theorems and the correspondence theorem in Section 8.1. Finally, in Section 9.3, central series are discussed and nilpotent groups are characterized as direct products of p -groups, and the Frattini and Fitting subgroups are introduced.

Chapter 10. Galois Theory. Galois groups of field extensions are defined, separable elements are introduced, and the main theorem of Galois theory is proved. Then it is shown that polynomials of degree 5 or more are not solvable in radicals. This requires only Chapter 6 (the reference to solvable groups in Section 10.3 is adequately reviewed there). Finally, cyclotomic polynomials are discussed and used (with the class equation) to prove Wedderburn’s theorem that every finite division ring is a field.

Chapter 11. Finiteness Conditions for Rings and Modules. The ascending and descending chain conditions on a module are introduced and the Jordan–Hölder theorem is proved. Then endomorphism rings are used to prove Wedderburn’s theorem that a simple, left artinian ring is a matrix ring over a division ring. Next, semisimple modules are studied and the results are employed to prove the Wedderburn–Artin theorem that a semisimple ring is a finite product of matrix rings over division rings. In addition, it is shown that these semisimple rings are characterized as the rings with every module projective and as the semiprime, left artinian rings.

Chapter Dependency Diagram



A Dashed arrow indicates minor dependency.

Acknowledgments

I express my appreciation to the following people for their useful comments and suggestions for the first edition of the book: F. Doyle Alexander, Stephen F. Austin State University; Steve Benson, Saint Olaf College; Paul M. Cook II, Furman University; Ronald H. Dalla, Eastern Washington University; Robert Fakler, University of Michigan–Dearborn; Robert M. Guralnick, University of Southern California; Edward K. Hinson, University of New Hampshire; Ron Hirschorn, Queen’s University; David L. Johnson, Lehigh University; William R. Nico, California State University–Hayward; Kimmo I. Rosenthal, Union College; Erik Shreiner (deceased), Western Michigan University; S. Thomeier, Memorial University; and Marie A. Vitulli, University of Oregon.

I also want to thank all the readers who informed me about typographical and other minor errors in the third edition. Particular thanks go to:

Carl Faith, Rutgers University, for giving the book a careful study and making many very useful suggestions, too numerous to list here;
David French, Derbyshire, UK, for pointing out several typographical errors;
Michel Racine, Université d’Ottawa, for pointing out a mistake in an exercise deducing the commutativity of addition in a ring from the other axioms;
Yoji Yoshii, Université d’Ottawa, for revealing two errors in the exercises for Chapter 5;
Yiqiang Zhou, Memorial University of Newfoundland, for many helpful suggestions and comments.

For the fourth edition, special thanks go to:

Jerome Lefebvre, University of Ottawa, for pointing out several typographical errors;

Edgar Goodaire and his students, Memorial University, for finding dozens of typographical errors and making many useful suggestions;

Keith Conrad, University of Connecticut, for many useful comments on the exposition;

Nazih Nahlus, American University of Beirut, for the proof that a finite multiplicative group of a field is cyclic;

Matthew Greenberg, University of Calgary, for pointing out that Burnside's lemma on Counting Orbits was due to Cauchy and Frobenius.

Milosz Kosmider, student, for correcting an error in Chapter 0;

Yannis Avrithis, National Technical University of Athens, for pointing out dozens of typographical errors and making several suggestions.

It is a pleasure to thank Steve Quigley for his generous assistance throughout the project. Thanks also go to the production staff at Wiley and particularly to Susanne Steitz-Filler for keeping the project on schedule and responding so quickly to all my questions. I also want to thank Joanne Canape for her vital assistance with the computer aspects of the project.

Finally, I want to thank my wife, Kathleen, for her unfailing support. Without her understanding and cooperation during the many hours that I was absorbed with this project, this book would not exist.

Notation used in the Text

Symbol	Description	First Used
\Rightarrow	implication	2
\Leftrightarrow	logical equivalence	3
\in	set membership	5
\subseteq	set containment	5
\subset	proper set containment	5
\mathbb{N}	set of natural numbers	5
\mathbb{Z}	set of integers	5
\mathbb{Q}	set of rational numbers	5
\mathbb{R}	set of real numbers	6
\mathbb{C}	set of complex numbers	6
$\mathbb{Z}^+, \mathbb{Q}^+, \mathbb{R}^+$	positive elements in these sets	6
\emptyset	empty set	6
\cup	union of sets	7
\cap	intersection of sets	7
$A \setminus B$	difference set	7
(a, b)	ordered pair	7
$A \times B$	cartesian product of sets A and B	8
(a_1, a_2, \dots, a_n)	ordered n -tuple	8
$\alpha : A \rightarrow B$	mapping α from A to B	10
$A \xrightarrow{\alpha} B$	image of x under mapping α	10
$\text{im}(\alpha)$	image of mapping α	12
$ A $	number of elements in set A	12
$\beta\alpha$	composite of mappings α and β	12
1_A	identity mapping on set A	13
α^{-1}	inverse of mapping α	14
\equiv	equivalence relation	17
$[a]$	equivalence class of a	17

Symbol	Description	First Used
A_{\equiv}	quotient set of equivalence \equiv	19
$n!$	n factorial	26
$\binom{n}{r}$	binomial coefficient	26
$d n$	d is a divisor of n	33
$\gcd(m, n)$		
$\gcd(n_1, \dots, n_r)$	greatest common divisor	33, 39
$\lcm(m, n)$		
$\lcm(n_1, \dots, n_r)$	least common multiple	39
$a \equiv b \pmod{n}$	congruence modulo n	43
\bar{a}	residue class of an integer a	43
\mathbb{Z}_n	integers modulo n	43
S_n	symmetric group of degree n	54
$(\begin{smallmatrix} 1 & \dots & n \\ \sigma_1 & \dots & \sigma_n \end{smallmatrix})$	permutation σ in S_n	54
ε	identity permutation in S_n	55
$(k_1 \ k_2 \ \dots \ k_r)$	cycle permutation in S_n	58
A_n	alternating group of degree n	62
$\operatorname{sgn} \sigma$	sign of permutation σ	66, 138
a^n	n th power of a	72
a^{-1}	inverse of a	73
\mathbb{C}^0	circle group	77
U_n	group of n th roots of unity	77
M^*	group of units of monoid M	79
S_X	group of permutations of set X	79
$GL_n(R)$	general linear group over R	80
C_n	cyclic group of order n	82
K_4	Klein 4-group	83
$SL_n(R)$	special linear group over R	138
$PSL_n(F)$	projective special linear group over F	398
$Z(G)$	center of group G	87
$\langle g \rangle$	cyclic subgroup generated by g	91
$o(g)$	order of group element g	92
$\langle X \rangle$	subgroup generated by X	96
$\operatorname{aut} G$	automorphism group of G	104
$\operatorname{inn} G$	inner automorphism group of G	105
Ha, aH	right, left cosets of subgroup H	109
$ G : H $	index of subgroup H in G	111
D_n	dihedral group	113
$H \triangleleft G$	H is a normal subgroup of G	122
Q	quaternion group	127
G/K	factor group of G by K	132
G'	derived (commutator) subgroup of G	134
$\ker \alpha$	kernel of a homomorphism α	137
B^n	set of binary n -tuples	144
$F(X, R)$	ring of functions $X \rightarrow R$	161
$M_n(R)$	ring of $n \times n$ matrices over R	161
$\operatorname{char} R$	characteristic of a ring R	163

Symbol	Description	First Used
$\mathbb{Z}(i)$	ring of gaussian integers	164
$T_2(R)$	upper triangular matrices over R	164
$Z(R)$	center of a ring R	165
R^{op}	opposite ring	169
\mathbb{H}	quaternions	174
$\text{ann}(a)$	annihilator of element a	182
R^1	ring extension of a general ring R	194
$R[x]$	ring of polynomials in x over R	203
$\deg f$	degree of polynomial f	205
$\Phi_n(x)$	cyclotomic polynomials	221
$a \sim b$	associates in an integral domain	253
$\text{span}\{v_1, \dots, v_n\}$	space spanned by v_1, \dots, v_n	277
$\dim V$	dimension of vector space V	279
$[E : F]$	dimension of E over a subfield F	283
$F(u_1, \dots, u_n)$	field generated over F by u_1, \dots, u_n	284
\mathbb{A}	field of algebraic numbers	289
f'	formal derivative of f	299
$GF(p^n)$	Galois field of order p^n	300
$N_1 \oplus N_2 \oplus \dots \oplus N_k$	direct sum of modules	325, 329
M^n	direct sum of n copies of module M	325
rank M	rank of free module M	333
$T(M)$	torsion submodule of M	334, 336
$M(p)$	p -primary component of M	337
class a	conjugacy class of a	358
$N(X)$	normalizer of a subgroup X	359
core H	core of a subgroup H	364
$G \cdot x$	orbit of x generated by G	367
$S(x)$	stabilizer of x	368
n_p	number of Sylow p -subgroups	374
$K \times_{\theta} H$	semidirect product of K by H	380
length G	composition length of G	390
$G^{(i)}$	higher derived subgroups for G	396
$Z_i(G), \Gamma_i(G)$	central series for group G	402, 403
$\Phi(G)$	Frattini subgroup of G	406
$F(G)$	Fitting subgroup of G	408
$\text{gal}(E : F)$	Galois group of E over F	413
K'	automorphisms fixing subfield K	425
H°	elements fixed by subgroup H	425
$S_R(x_1, x_2, \dots, x_n)$	elementary symmetric polynomials	439
$\mathbb{Z}_{p^{\infty}}$	The Prüfer group for a prime p	449
$\text{hom}(M, N)$	group of module homomorphisms	452
$\text{end}(M)$	endomorphism ring of module M	452
$H(K)$	homogeneous component	461
$\text{re } z, \text{im } z$	real, imaginary, part of z	472
$\bar{z}, z $	conjugate, absolute value of z	473
$e^{i\theta}$	notation for $\cos \theta + i \sin \theta$	374

A Sketch of the History of Algebra to 1929

- 2500 BC Hieroglyphic numerals used in Egypt.
- 2400 BC Babylonians begin positional algebraic notation.
- 600 BC Pythagoreans discuss prime numbers.
- 250 Diophantus writes *Arithmetica*, using notation from which modern notation evolved, and insists on exact solutions of equations in integers.
- 830 al-Khowarizmi writes *Al-jabr*, a textbook giving rules for solving linear and quadratic equations.
- 1202 Leonardo of Pisa writes *Liber abaci* on arithmetic and algebraic equations.
- 1545 Tartaglia solves the cubic, and Cardano publishes the result in his *Ars Magna*. Imaginary numbers are suggested.
- 1580 Viète uses vowels to represent unknown quantities, with consonants for constants.
- 1629 Fermat becomes the founder of the modern theory of numbers.
- 1636 Fermat and Descartes invent analytic geometry, using algebra in geometry.
- 1749 Euler formulates the fundamental theorem of algebra.
- 1771 Lagrange solves the general cubic and quartic by considering permutations of the roots.
- 1799 Gauss publishes his first proof of the fundamental theorem of algebra.
- 1801 Gauss publishes his *Disquisitiones Arithmeticae*.
- 1813 Ruffini claims that the general quintic cannot be solved by radicals.
- 1824 Abel proves that the general quintic cannot be solved by radicals.
- 1829 Galois introduces groups of substitutions.
- 1831 Galois sends his great memoir to the French Académie, but it is rejected.
- 1843 Hamilton discovers the quaternions.

- 1846 Kummer invents his ideal numbers.
- 1854 Cayley introduces the multiplication table of a group.
- 1870 Jordan publishes his monumental *Traité*, which explains Galois theory, develops group theory, and introduces composition series.
- 1870 Kronecker proves the fundamental theorem of finite abelian groups.
- 1872 Sylow presents his results on what are now called the Sylow theorems.
- 1878 Cayley proves that every finite group can be represented as a group of permutations.
- 1879 Dedekind defines algebraic number fields, studies the factorization of algebraic integers into primes, and introduces the concept of an ideal.
- 1889 Peano formulates his axioms for the natural numbers.
- 1889 Hölder completes the proof of the Jordan–Hölder theorem.
- 1905 Wedderburn proves that finite division rings are commutative.
- 1908 Wedderburn proves his structure theorem for finite dimensional algebras with no nilpotent ideals.
- 1921 Noether publishes her influential paper on chain conditions in ring theory.
- 1927 Artin extends Wedderburn's 1908 paper to rings with the descending chain condition.
- 1929 Noether establishes the modern approach to the theory of representations of finite groups.

Chapter 0

Preliminaries

The science of Pure Mathematics, in its modern development, may claim to be the most original creation of the human spirit.

—Alfred North Whitehead

This brief chapter contains background material needed in the study of abstract algebra and introduces terms and notations used throughout the book. Presenting all this information at the beginning is preferable, because its introduction at the point it is needed interrupts the continuity of the text. Moreover, we can include enough detail here to help those readers who may be less prepared or are using the book for self-study. However, much of this material may be familiar. If so, just glance through it quickly and begin with Chapter 1, referring to this chapter only when necessary.

0.1 PROOFS

The essential quality of a proof is to compel belief.

—Pierre de Fermat

Logic plays a basic role in human affairs. Scientists use logic to draw conclusions from experiments, judges use it to deduce consequences of the law, and mathematicians use it to prove theorems. Logic arises in ordinary speech with assertions such as “if John studies hard, he will pass the course,” or “if an integer n is divisible by 6, then n is divisible by 3.” In each case, the aim is to assert that if a certain statement is true, then another statement must also be true. In fact, if p and q

denote statements, most theorems take the form of an **implication**: “If p is true, then q is true.” We write this in symbols as

$$p \Rightarrow q$$

and read it as “ p implies q .” Here, p is the **hypothesis** and q the **conclusion** of the implication. Verification that $p \Rightarrow q$ is valid is the **proof** of the implication. In this section, we examine the most common methods of proof¹ and illustrate each technique with an example.

Method of Direct Proof. To prove $p \Rightarrow q$, demonstrate directly that q is true whenever p is true.

Example 1. If n is an odd integer, show that n^2 is odd.

Solution. If n is odd, it has the form $n = 2k + 1$ for some integer k . Then $n^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$ is also odd because $2k^2 + 2k$ is an integer. \square

Note that the computation $n^2 = 4k^2 + 4k + 1$ in Example 1 involves some simple properties of arithmetic that we did not prove. Actually, a whole body of mathematical information lies behind nearly every proof of any complexity, although this fact usually is not stated explicitly.

Suppose that you are asked to verify that $n^2 \geq 0$ for every integer n . This expression is an implication: If n is an integer, then $n^2 \geq 0$. To prove it, you might consider separately the cases that $n > 0$, $n = 0$, and $n < 0$ and then show that $n^2 \geq 0$ in each case. (You would have to invoke the fact that $0^2 = 0$ and that the product of two positive, or two negative, integers is positive.) We formulate the general method as follows:

Method of Reduction to Cases. To prove $p \Rightarrow q$, show that p implies at least one of a list p_1, p_2, \dots, p_n of statements (the cases) and that $p_i \Rightarrow q$ for each i .

Example 2. If n is an integer, show that $n^2 - n$ is even.

Solution. Note that $n^2 - n = n(n - 1)$ is even if n or $n - 1$ is even. Hence, given n , we consider the two cases that n is even or odd. Because $n - 1$ is even in the second case, $n^2 - n$ is even in either case. \square

The statements used in mathematics must be true or false. This requirement leads to a proof technique that can mystify beginners. The method is a formal version of a debating strategy whereby the debater assumes the truth of an opponent’s position and shows that it leads to an absurd conclusion.

Method of Proof by Contradiction. To prove $p \Rightarrow q$, show that the assumption that both p is true and q is false leads to a contradiction.

Example 3. If r is a rational number (fraction), show that $r^2 \neq 2$.

Solution. To argue by contradiction, we assume that r is a rational number and that $r^2 = 2$ and show that this assumption leads to a contradiction. Let m and n be integers such that $r = \frac{m}{n}$ is in lowest terms (so, in particular, m and n are both not even). Then $r^2 = 2$ gives $m^2 = 2n^2$, so m^2 is even. This means m is even

¹For a more detailed look at proof techniques, see Solow, D., *How to Read and Do Proofs*, 2nd ed., Wiley, 1990; Lucas, J.F., *Introduction to Abstract Mathematics*, Wadsworth, 1986, Chapter 2.

(Example 1), say $m = 2k$. But then $2n^2 = m^2 = 4k^2$, so $n^2 = 2k^2$ is even, and hence n is even. This shows that n and m are both even, contrary to the choice of n and m . \square

Example 4. If $2^n - 1$ is a prime number, show that n is a prime number. (Here, a prime number is an integer greater than 1 that cannot be factored as the product of two smaller positive integers.)

Solution. We must show that $p \Rightarrow q$, where p is “ $2^n - 1$ is a prime” and q is “ n is a prime.” Suppose that q is false so that n is not a prime, say $n = ab$, where $a \geq 2$ and $b \geq 2$ are integers. If we write $2^a = x$, then $2^n = 2^{ab} = (2^a)^b = x^b$. Hence,

$$2^n - 1 = x^b - 1 = (x - 1)(x^{b-1} + x^{b-2} + \cdots + x^2 + x + 1).$$

As $x \geq 4$, this factors $2^n - 1$ into smaller positive integers, a contradiction. \square

The next example exhibits one way to show that an implication is *not* valid.

Example 5. Show that the implication “ n is a prime $\Rightarrow 2^n - 1$ is a prime” is false.

Solution. The first few primes are $n = 2, 3, 5, 7$, and the corresponding values $2^n - 1 = 3, 7, 31, 127$ are all prime, as the reader can verify. This observation seems to be evidence that the implication is true. However, the next prime is $n = 11$ and $2^{11} - 1 = 2047 = 23 \cdot 89$ clearly is not a prime. \square

We say that $n = 11$ is a **counterexample** to the (proposed) implication in Example 5. Note that if you can find even one example for which an implication is not valid, the implication is false. Thus, disproving implications in a sense is easier than proving them.

The implications in Examples 4 and 5 are closely related: They have the form $p \Rightarrow q$ and $q \Rightarrow p$, where p and q are statements. Each is called the **converse** of the other, and as the examples show, an implication can be valid even though its converse is not valid. If both $p \Rightarrow q$ and $q \Rightarrow p$ are valid, the statements p and q are called **logically equivalent**, which we write in symbols as

$$p \Leftrightarrow q$$

and read “ p if and only if q .” Many of the most satisfying theorems make the assertion that two statements, ostensibly quite different, are in fact logically equivalent.

Example 6. If n is an integer, show that “ n is odd $\Leftrightarrow n^2$ is odd.”

Solution. In Example 1, we proved the implication “ n is odd $\Rightarrow n^2$ is odd.” Here, we prove the converse by contradiction. If n^2 is odd, we assume that n is not odd. Then n is even, say $n = 2k$, so $n^2 = 4k^2$ is also even, a contradiction. \square

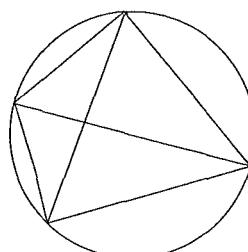
Many more examples of proofs can be found in this book and, although they are often more complex, most are based on one of these methods. In fact, abstract algebra is one of the best topics on which the reader can sharpen his or her skill at constructing proofs. Part of the reason for this is that much of abstract algebra is developed using the **axiomatic method**. That is, in the course of studying various examples, it is observed that they all have certain properties in common. Then when a general abstract system is studied in which these properties are *assumed* to hold (and are called **axioms**), statements (called **theorems**) are deduced from these

axioms by using the methods presented in this section. These theorems will then be true in *all* the concrete examples because the axioms hold in each case. But this procedure is more than just an efficient method for finding theorems in examples. By reducing the proof to its essentials, we gain a better understanding of why the theorem is true and how it relates to analogous theorems in other abstract systems.

The axiomatic method is not new. Euclid first used it in about 300 BC to derive all the propositions of (euclidean) geometry from a list of 10 axioms. The method lends itself well to abstract algebra. The axioms are simple and easy to understand, and there are only a few of them. For example, group theory contains a large number of theorems derived from only four simple axioms.

Exercises 0.1

1. In each case, prove the result and either prove the converse or give a counterexample.
 - (a) If n is an even integer, then n^2 is a multiple of 4.
 - (b) If m is an even integer and n is an odd integer, then $m + n$ is odd.
 - (c) If $x = 2$ or $x = 3$, then $x^3 - 6x^2 + 11x - 6 = 0$.
 - (d) If $x^2 - 5x + 6 = 0$, then $x = 2$ or $x = 3$.
2. In each case, prove the result by splitting into cases or give a counterexample.
 - (a) If n is any integer, then $n^2 = 4k + 1$ for some integer k .
 - (b) If n is any odd integer, then $n^2 = 8k + 1$ for some integer k .
 - (c) If n is any integer, $n^3 - n = 3k$ for some integer k . [Hint: Use the fact that each integer has one of the forms $3k$, $3k + 1$, or $3k + 2$, where k is an integer.]
3. In each case, prove the result by contradiction and either prove the converse or give a counterexample.
 - (a) If $n > 2$ is a prime integer, then n is odd.
 - (b) If $n + m = 25$, where n and m are integers, one of n and m is greater than 12.
 - (c) If a and b are positive numbers and $a \leq b$, then $\sqrt{a} \leq \sqrt{b}$.
 - (d) If m and n are integers and mn is even, then m is even or n is even.
4. Prove each implication by contradiction.
 - (a) If x and y are positive numbers, then $\sqrt{x+y} \neq \sqrt{x} + \sqrt{y}$.
 - (b) If x is irrational and y is rational, then $x+y$ is irrational.
 - (c) If 13 people are selected, at least 2 have birthdays in the same month.
 - (d) **Pigeonhole Principle.** If $n+1$ pigeons are placed in n holes, some hole contains at least two pigeons.
5. Disprove each statement by giving a counterexample.
 - (a) $n^2 + n + 11$ is a prime for all positive integers n .
 - (b) $n^3 \geq 2^n$ for all integers $n \geq 2$.
 - (c) If n points are arranged on a circle in such a way that no three of the lines joining them have a common point, these lines divide the circle into 2^{n-1} regions. For example, if $n = 4$, there are $8 = 2^3$ regions as shown in the figure.



6. If p is a statement, let $\sim p$ denote the statement “not p ,” called the **negation** of p . Thus, $\sim p$ is true when p is false, and false when p is true. Show that if $\sim q \Rightarrow \sim p$, then $p \Rightarrow q$. [The implication $\sim q \Rightarrow \sim p$ is called the **contrapositive** of $p \Rightarrow q$.]

0.2 SETS

No one shall expel us out of the paradise which Cantor has created for us.

—David Hilbert

Everyone has an idea of what a set is. If asked to define it, you would likely say that “a set is a collection of objects” or something similar. However, such a response just shifts the question to what a collection is, without any gain at all. To add to the problem, when you think of concrete examples of sets, such as the set of all atoms in the earth, or even of more abstract examples, such as the set of all positive integers, you can see at once that the idea of a set is closely related to another idea, that of *membership* in a set. These ideas are so fundamental that we make no attempt to define them, taking them as primitive concepts in the theory of sets. We then use them to define the other concepts of the theory intuitively. Certain basic properties of sets must be assumed (the axioms of the theory), but it is not our intention to pursue this axiomatic development here. Instead, we rely on intuitive ideas about sets to enable us to describe enough of set theory to provide the language of abstract algebra.

Hence, we consider **sets** and call the members of a set the **elements** of the set. Sets are usually denoted by uppercase letters and elements by lowercase letters. The fact that a is an element of set A is denoted

$$a \in A.$$

If A and B are sets, we say that A is **contained** in B if every element of A is an element of B . In this case, we say that A is a **subset** of B and write

$$A \subseteq B \quad \text{or equivalently} \quad B \supseteq A.$$

The intuitive idea that two sets are the same if they have the same elements is reflected in the following axiom.

Principle of Set Equality. *If A and B are sets, then*

$$A = B \quad \text{if and only if} \quad A \subseteq B \quad \text{and} \quad B \subseteq A.$$

This principle is useful because often the easiest way to show that $A = B$ is to verify separately that $A \subseteq B$ and $B \subseteq A$. We use it frequently, often without comment.

If it is not the case that $A = B$, we write $A \neq B$. Similarly, we frequently use the notations $x \notin A$ and $A \not\subseteq B$. If $A \subseteq B$ but $A \neq B$, we write $A \subset B$ and refer to A as a **proper subset** of B .

Several important sets of numbers are represented by special symbols:

\mathbb{N} —the **set of natural numbers** (*positive integers and zero*)

\mathbb{Z} —the **set of integers** (*whole numbers, positive, negative, and zero*)

\mathbb{Q} —the **set of rational numbers** (*quotients $\frac{m}{n}$ of integers, where $n \neq 0$*)

\mathbb{R} —the set of real numbers

\mathbb{C} —the set of complex numbers

These notations are used throughout the book. Note that $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$. We write \mathbb{Z}^+ , \mathbb{Q}^+ , and \mathbb{R}^+ for the set of positive elements in these sets.

The only way to completely describe a set is to specify its elements in some unambiguous way. If the set has a finite number of elements, this is often accomplished by listing the elements. Thus, we can describe the set A of positive integers that are less than 6 as

$$A = \{1, 2, 3, 4, 5\}.$$

We frequently describe the elements in a set as those members of some known set that have a certain property. Thus, the set A may be described as follows:

$$A = \{x \in \mathbb{Z} \mid 1 \leq x \leq 5\},$$

which we read as “the set of elements x in \mathbb{Z} such that $1 \leq x \leq 5$.” More generally, if $p(x)$ is any statement about the elements x of a known set U , the set of all elements x of U for which $p(x)$ is true is denoted

$$\{x \in U \mid p(x)\}.$$

This notation has some variations, such as

$$\begin{aligned} \{0, 3, 6\} &= \{x \in \mathbb{Z} \mid x \text{ is a multiple of } 3 \text{ and } 0 \leq x \leq 6\} \\ &= \{x \in \mathbb{R} \mid x^3 - 9x^2 + 18x = 0\} \\ &= \{3x \mid x = 0, 1, 2\}. \end{aligned}$$

We use such notations without further comment.

If a finite set A has n elements, we often denote A as

$$A = \{a_1, a_2, \dots, a_n\} = \{a_i \mid 1 \leq i \leq n\}.$$

We denote the number of elements in a finite set A as $|A|$ and call sets with $|A| = 1$ **singletons**. If a set A is not finite, we say that A is **infinite** and write $|A| = \infty$. Sometimes we list infinite sets; for example, $B = \{3, 5, 7, \dots\}$ indicates the set of odd integers greater than 1. However, this notation can be ambiguous; for example, B could indicate the set of odd primes. Actually,

$$B = \{2k + 1 \mid k \in \mathbb{Z}, k \geq 1\}$$

is a much better description of B because it reveals the pattern used to describe the elements. Nonetheless, we use descriptions such as $B = \{3, 5, 7, \dots\}$ when the meaning is clear from the context.

We assume (this is an axiom) that there exists a set with *no* elements. This set is called the **empty set** and is denoted \emptyset . Thus, $\{x \mid x \in \mathbb{R}, x^2 = -1\} = \emptyset$ because there is *no* real number x with $x^2 = -1$. The following property of \emptyset is used frequently:

$$\emptyset \subseteq A \text{ for every set } A.$$

The verification of this assertion provides a nice example of proof by contradiction. Observe that $\emptyset \not\subseteq A$ implies the existence of an element $x \in \emptyset$ such that $x \notin A$; a contradiction since \emptyset has no element.

Let A_1, A_2, \dots, A_n be sets. We define their **union** $A_1 \cup A_2 \cup \dots \cup A_n$ and their **intersection** $A_1 \cap A_2 \cap \dots \cap A_n$ as follows:

$$A_1 \cup A_2 \cup \dots \cup A_n = \{x \mid x \in A_i \text{ for some } i = 1, 2, \dots, n\},$$

$$A_1 \cap A_2 \cap \dots \cap A_n = \{x \mid x \in A_i \text{ for every } i = 1, 2, \dots, n\}.$$

These sets sometimes are denoted $\cup_{i=1}^n A_i$ and $\cap_{i=1}^n A_i$, respectively.

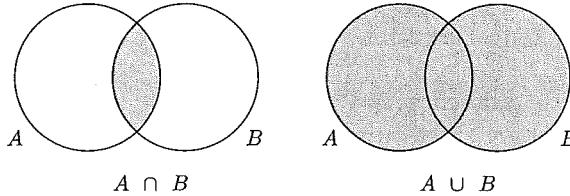
The intersection $A_1 \cap A_2 \cap \dots \cap A_n$ is a subset of each of the sets A_i , and it contains every such subset. Similarly, the union $A_1 \cup A_2 \cup \dots \cup A_n$ contains each of the sets A_i and is contained in every such set.

If only two sets A and B are involved, we have

$$A \cup B = \{x \mid x \in A \text{ or } x \in B, \text{ or both}\},$$

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

The use of *Venn diagrams*, named after the English logician John Venn, clarifies many properties of these operations. Points inside some region of the plane (say, the interior of a circle) represent the elements of a set. Then the shaded regions in the diagram represent the sets $A \cap B$ and $A \cup B$.



Using the principle of set equality, the following properties can be proved for arbitrary sets A , B , and C :

$$\begin{array}{lll} A \cup A = A & A \cup B = B \cup A & A \cup (B \cup C) = (A \cup B) \cup C, \\ A \cap A = A & A \cap B = B \cap A & A \cap (B \cap C) = (A \cap B) \cap C. \end{array}$$

These are called the **idempotent**, **commutative**, and **associative** laws, respectively. In addition, we have the **distributive** laws

$$\begin{aligned} A \cup (B_1 \cap B_2 \cap \dots \cap B_n) &= (A \cup B_1) \cap (A \cup B_2) \cap \dots \cap (A \cup B_n), \\ A \cap (B_1 \cup B_2 \cup \dots \cup B_n) &= (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n). \end{aligned}$$

The **difference** $A \setminus B$ of two sets consists of the elements of A that are not in B , more formally

$$A \setminus B = \{x \mid x \in A \text{ and } x \notin B\}.$$

This notation arises frequently, primarily for descriptive purposes.

The sets $\{a, b\}$ and $\{b, a\}$ are equal because the order in which the elements of a set are listed is irrelevant. However, taking the order into consideration is frequently useful. A pair of elements is called an **ordered pair** when they are taken to be in a definite order. The notation

$$(a, b)$$

denotes the ordered pair in which the first member is a and the second is b . The defining property is

$$(a, b) = (a_1, b_1) \quad \text{if and only if} \quad a = a_1 \quad \text{and} \quad b = b_1.$$

Thus, a and b are uniquely determined by the ordered pair (a, b) , and they are called the first and second **components** of the ordered pair. In particular, (a, b) and (b, a) are *distinct* ordered pairs (assuming that $a \neq b$), in contrast to the *equal* sets $\{a, b\}$ and $\{b, a\}$. The most familiar use of ordered pairs is in describing the coordinates (x, y) of a point in the euclidean plane.

The **cartesian product** $A \times B$ of two sets A and B is defined to be the set

$$A \times B = \{(a, b) \mid a \in A, b \in B\}$$

of all ordered pairs with the first component from A and the second component from B .

The sets A and B can be equal here, and $A \times A$ is sometimes expressed as A^2 . For example, if $A = \{1, 2\}$ and $B = \{1, 2, 3\}$,

$$\begin{aligned} A \times A &= A^2 = \{(1, 1), (1, 2), (2, 1), (2, 2)\}, \\ A \times B &= \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3)\}. \end{aligned}$$

Clearly, $\mathbb{R} \times \mathbb{R}$ is the euclidean plane, and this is the source of the term *cartesian*. The name honors René Descartes, who used such coordinates in his work on geometry.²

By analogy with ordered pairs, we call a set of elements a_1, a_2, \dots, a_n an **ordered n -tuple** if they are arranged in a definite order. We use the notation

$$(a_1, a_2, \dots, a_n)$$

for ordered n -tuples, and the defining property is

$$(a_1, a_2, \dots, a_n) = (b_1, b_2, \dots, b_n) \quad \text{if and only if} \quad a_i = b_i \quad \text{for each } i.$$

We call a_i the *i th component* of the n -tuple (a_1, a_2, \dots, a_n) . If A_1, A_2, \dots, A_n are sets, their cartesian product $A_1 \times A_2 \times \dots \times A_n$ is defined to be the set

$$A_1 \times A_2 \times \dots \times A_n = \{(a_1, a_2, \dots, a_n) \mid a_i \in A_i \text{ for each } i\}$$

of all ordered n -tuples whose i th component belongs to A_i for each i .

Exercises 0.2

1. In each case, describe A in the notation $A = \{x \mid p(x)\}$.
 - (a) A is the set of all positive multiples of 5.
 - (b) A is the set of all integers between $-\frac{1}{2}$ and $\frac{9}{2}$.
2. List the elements of the following sets.

(a) $\{n \in \mathbb{N} \mid n^3 \text{ is odd}\}$ (c) $\{x \in \mathbb{R} \mid x^3 + 3x^2 - x - 3 = 0\}$ (e) $\{x \in \mathbb{Q} \mid x^2 = 2\}$	(b) $\{n \in \mathbb{N} \mid 2n + 1 < 16\}$ (d) $\{\frac{1}{n^2} \mid n \in \mathbb{Z}, n \neq 0\}$ (f) $\{n \in \mathbb{N} \mid 2 < 3n + 1 < 20\}$
---	---

²Actually these coordinates were known and used much earlier by Nicole Oresme (1323–1382). See Boyer, C.B., *A History of Mathematics*, New York: Wiley, 1968, p. 379.

3. Which of the following pairs of sets are equal? Defend your answer.
- | | |
|---|--|
| (a) $A = \{n \in \mathbb{Z} \mid n^2 \leq 4\}$ | $B = \{x \in \mathbb{R} \mid x^2 - 3x + 2 = 0\}$ |
| (b) $A = \{n \in \mathbb{Z} \mid n = \frac{1}{n}\}$ | $B = \{x \in \mathbb{R} \mid x^2 = 1\}$ |
| (c) $A = \text{the set of letters in "alloy"}$ | $B = \text{the set of letters in "loyal"}$ |
| (d) $A = \{2, \{3\}, 4\}$ | $B = \{2, \{3, 4\}\}$ |
| (e) $A = \{1\}$ | $B = \{\{1\}\}$ |
| (f) $A = \{x \in \mathbb{R} \mid x^2 = -1\}$ | $B = \{x \in \mathbb{Q} \mid x^2 = 2\}$ |
| (g) $A = \{x \in \mathbb{Z} \mid x^2 \leq 1\}$ | $B = \{x \in \mathbb{R} \mid x^3 = x\}$ |
4. Let $A = \{1, 2, 3, 4\}$, $B = \{1, 2, 3\}$, and $C = \{2, 4\}$. Find all sets X satisfying each pair of conditions.
- | | |
|---|---|
| (a) $X \subseteq B$ and $X \subseteq C$ | (b) $X \subseteq A$ and $X \not\subseteq B$ |
| (c) $X \subseteq B$ and $X \not\subseteq C$ | (d) $X \subset B$ and $X \not\subseteq C$ |
5. In each case, prove the assertion if it is true or give a counterexample if it is false.
(We temporarily suspend the convention of denoting elements by lowercase letters.)
- | | |
|---|---|
| (a) If $A \in B$ and $B \subseteq C$, then $A \in C$. | (b) If $A \in B$ and $B \in C$, then $A \in C$. |
| (c) If $A \in B$ and $B \subseteq C$, then $A \subseteq C$. | (d) If $A \subseteq B$ and $B \in C$, then $A \in C$. |
6. (a) Show that $A \cap B$ is the largest common subset of A and B in the sense that it contains every such common subset.
(b) Show that $A \cup B$ is the smallest set containing both A and B in the sense that it is contained in every such set.
7. Prove the distributive laws using the principle of set equality.
8. Let A and B be sets. If $A \cap X = B \cap X$ and $A \cup X = B \cup X$ for some set X , prove that $A = B$. [Hint: $A = A \cap (A \cup X)$.]
9. Find sets A , B , and C such that $A \cap B \cap C = \emptyset$ but that none of $A \cap B$, $A \cap C$, and $B \cap C$ is empty.
10. (a) If A and B are nonempty sets and $A \times B = B \times A$, show that $A = B$.
(b) Show that $A \times B = B \times A$ if and only if either $A = B$ or one of A and B is empty.
(c) Show that $A \cap B = \{x \mid (x, x) \in A \times B\}$.
11. (a) Prove that $A \times (B \cap C) = (A \times B) \cap (A \times C)$.
(b) Prove that $A \times (B \cup C) = (A \times B) \cup (A \times C)$.
(c) Prove that $(A \cap B) \times (A' \cap B') = (A \times A') \cap (B \times B')$.
12. Care must be taken in defining sets. Consider

$$R = \{X \mid X \text{ is a set and } X \text{ is not an element of itself}\}.$$

Show that R cannot be a set. [Hint: If R is a set, is R a member of itself or not?]

The assumption that R is a set is called the **Russell Paradox**, after Bertrand Russell.

0.3 MAPPINGS

The concept of a function is basic to all mathematics and real-valued functions are essential in calculus and elementary algebra. In this section, we introduce functions from any set A to any set B . These more general functions are called mappings to avoid confusion. In this generality, sets and mappings are the language of abstract algebra.

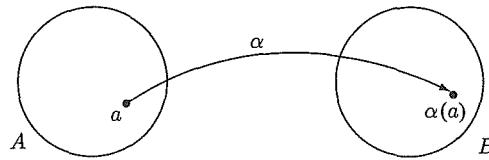
In many applications of set theory, we are interested in some property or attribute of the elements a of a set A . For example, if A is the set of all people, the attribute of $a \in A$ might be the age of a or the gender of a . In each case, the attribute is itself an element of another set B (in the latter case, $B = \{F, M\}$)

will do). Hence, for each $a \in A$, there is a uniquely determined attribute $b \in B$. The assignment $a \mapsto b$ is an example of a mapping.³ In general, if A and B are sets, a **mapping** (or **function**) α from A to B , written

$$\alpha : A \rightarrow B \quad \text{or} \quad A \xrightarrow{\alpha} B,$$

is a rule⁴ that assigns to every element a of A exactly one element $\alpha(a)$ of B . This assignment is sometimes denoted $a \mapsto \alpha(a)$ (see the diagram). We refer to A and B as the **domain** and **codomain**, respectively, of the mapping α . For $a \in A$, the unique element $b \in B$, such that $b = \alpha(a)$ is called the **image** of a under α . The notion of a mapping is one of the most fertile ideas in mathematics.

The process of defining a mapping α consists of two parts: First, we must specify the domain A and codomain B of α , and then, for every $a \in A$, we must specify exactly one element $\alpha(a)$ in B that α assigns to a . We refer to this latter task as defining the **action** of α and then say that the mapping is **well defined**. This can be done in several ways.



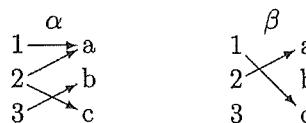
If the domain and codomain are sets of numbers, the most common way to define a mapping is by means of a formula. Thus, $\alpha(x) = x^2 + 1$ and $\beta(x) = 3x - 2$ define mappings $\mathbb{R} \rightarrow \mathbb{R}$. Sometimes the mapping is given by a different formula on different parts of the domain. For example,

$$\alpha : \mathbb{Z} \rightarrow \{1, -1\} \quad \text{given by} \quad \alpha(n) = \begin{cases} 1 & \text{if } n \text{ is even} \\ -1 & \text{if } n \text{ is odd} \end{cases}$$

is a mapping. We can describe mappings with a finite domain by simply listing the images of the domain's elements. For example, we can define $\alpha : \{1, 2, 3\} \rightarrow \{a, b, c\}$ by stipulating that $\alpha(1) = a$, $\alpha(2) = a$, and $\alpha(3) = c$. We describe this action graphically with an arrow diagram:



Example 1. Consider the correspondences α and β from $\{1, 2, 3\}$ to $\{a, b, c\}$ with actions given by the arrow diagrams:



³We will usually denote mappings by lowercase Greek letters $\alpha, \beta, \gamma, \dots$

⁴This definition has the difficulty that “rule” is just a synonym for “mapping.” This is circumvented by the **formal definition**: A mapping $\alpha : A \rightarrow B$ is a set $\alpha \subseteq A \times B$ of ordered pairs in which every element of A occurs exactly once as the first component of a pair in α . Then, for $a \in A$, the unique element $b \in B$ such that $(a, b) \in \alpha$ is denoted $b = \alpha(a)$.

Then α is not well defined because α assigns *both* a and c to 2, and β is not well defined because β assigns *no* element to 3.

Example 2. Let $\alpha : \mathbb{Q} \rightarrow \mathbb{Z}$ be given by $\alpha(\frac{n}{m}) = n$. Then α is not well defined. In fact, let $x = \frac{1}{2} = \frac{2}{4}$. Then $\alpha(x) = \alpha(\frac{1}{2}) = 1$ and $\alpha(x) = \alpha(\frac{2}{4}) = 2$, so the element of \mathbb{Z} assigned to x is not uniquely determined.

Two mappings are equal if and only if they have the same action.

Theorem 1. If $\alpha : A \rightarrow B$ and $\beta : A \rightarrow B$ are mappings, then

$$\alpha = \beta \quad \text{if and only if} \quad \alpha(a) = \beta(a) \text{ for all } a \in A.$$

Proof. The formal definition presents α and β as sets of ordered pairs: $\alpha = \{(a, \alpha(a)) \mid a \in A\}$ and $\beta = \{(a, \beta(a)) \mid a \in A\}$. Now Theorem 1 follows from the principle of set equality. ■

Example 3. Show that $\alpha = \beta$, where $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ and $\beta : \mathbb{R} \rightarrow \mathbb{R}$ are given for all $x \in \mathbb{R}$ by

$$\alpha(x) = x^2 + x + 1 \quad \text{and} \quad \beta(x) = (x - 1)(x + 2) + 3.$$

Solution. The fact that $x^2 + x + 1 = (x - 1)(x + 2) + 3$ is an **identity** in x (that is, it is true for all $x \in \mathbb{R}$) implies that $\alpha = \beta$. Such identities are the basis of many of the manipulations of mappings defined by formulas. □

One-to-One and Onto Mappings

Let $\alpha : A \rightarrow B$ be a mapping. For convenience, let us say that an element $b \in B$ is “hit” by α if $b = \alpha(a)$ for some $a \in A$, that is, if b is the image of some a in A . We say that α is **one-to-one** (or **injective**) if no element of B is “hit” more than once, that is, if (for a and a_1 in A)

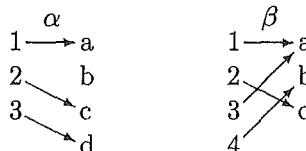
$$\alpha(a) = \alpha(a_1) \quad \text{implies} \quad a = a_1.$$

We say that α is **onto** (or **surjective**) if every element of B is “hit” at least once, that is,

Every $b \in B$ has the form $b = \alpha(a)$ for some $a \in A$.

A mapping that is both one-to-one and onto is called a **bijection** and is said to be **bijective**.

These notions are best illustrated by arrow diagrams. Consider the mappings $\alpha : \{1, 2, 3\} \rightarrow \{a, b, c, d\}$ and $\beta : \{1, 2, 3, 4\} \rightarrow \{a, b, c\}$ with the following actions:



Then α is one-to-one (no element is “hit” twice) but not onto (b is not “hit”), whereas β is onto (every element of $\{a, b, c\}$ is “hit”) but not one-to-one (a is “hit” twice).

Example 4. If $\alpha : \mathbb{N} \rightarrow \mathbb{N}$ is defined by $\alpha(n) = 2n + 1$ for all $n \in \mathbb{N}$, show that α is one-to-one but not onto.

Solution. If $\alpha(n) = \alpha(m)$, then $2n + 1 = 2m + 1$, whence $n = m$. This shows that α is one-to-one. But α is not onto because no even integer has the form $\alpha(n) = 2n + 1$ for $n \in \mathbb{N}$. \square

Example 5. Show that $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ given by $\alpha(x) = 2x - 5$ is a bijection.

Solution. If $\alpha(x) = \alpha(x_1)$, then $2x - 5 = 2x_1 - 5$. This implies that $x = x_1$, so α is one-to-one. To show that α is onto, we must demonstrate that each element $y \in \mathbb{R}$ (the codomain) has the form $y = \alpha(x)$ for some x in \mathbb{R} . This requirement is $y = 2x - 5$, which has a solution $x = \frac{1}{2}(y + 5)$ in \mathbb{R} for each y . \square

If $\alpha : A \rightarrow B$ is a mapping, the **image** of α is the set

$$\text{im}(\alpha) = \alpha(A) = \{\alpha(a) \mid a \in A\}$$

of all images of elements of A . Thus, $\alpha(A) \subseteq B$, and α is onto if and only if $\alpha(A) = B$. It is convenient sometimes to regard $\alpha : A \rightarrow \alpha(A)$. With this smaller codomain, it is clear that α is onto.

If $\alpha : A \rightarrow B$ is a bijection, the correspondence $a \leftrightarrow \alpha(a)$ pairs every element in each of the sets A and B with exactly one element of the other set. In particular, if both A and B are finite, they have the same number of elements. We write this as $|A| = |B|$, where $|X|$ denotes the number of elements in the finite set X .

We have presented examples of mappings that are onto and not one-to-one and mappings that are one-to-one and not onto. Theorem 2 covers an important situation in which these properties are equivalent.

Theorem 2. Let $\alpha : A \rightarrow B$ be a mapping where A and B are nonempty finite sets with $|A| = |B|$. Then α is one-to-one if and only if α is onto.

Proof. If α is one-to-one, then $\alpha : A \rightarrow \alpha(A)$ is a bijection, so $|A| = |\alpha(A)|$. Hence, $|\alpha(A)| = |B|$ and it follows that $\alpha(A) = B$ because $\alpha(A) \subseteq B$ and both sets are finite. This means that α is onto.

Conversely, let $|A| = |B| = n$ and write $B = \{b_1, b_2, \dots, b_n\}$, where b_i are distinct. Let $A_i = \{a \in A \mid \alpha(a) = b_i\}$ for each i . Then $A = A_1 \cup A_2 \cup \dots \cup A_n$, and $A_i \cap A_j = \emptyset$ whenever $i \neq j$ because b_i are distinct. It follows that $n = |A| = |A_1| + |A_2| + \dots + |A_n|$. But $|A_i| \geq 1$ for each i (because α is onto), so $|A_i| = 1$ for each i . This implies that α is one-to-one. \blacksquare

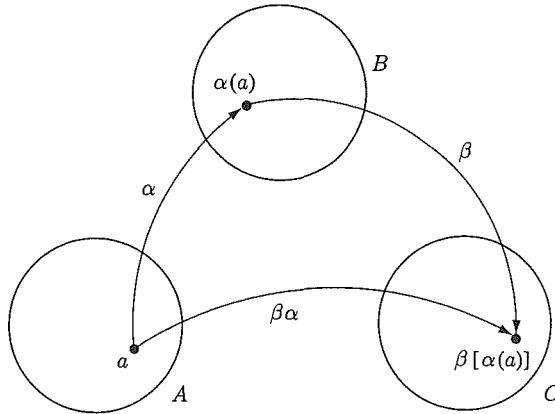
Composition and Inverse

Two linked mappings $A \xrightarrow{\alpha} B \xrightarrow{\beta} C$ may be combined naturally to obtain a mapping $A \rightarrow C$. In this case, we define the **composite** mapping

$$\beta\alpha : A \rightarrow C \quad \text{by} \quad \beta\alpha(a) = \beta[\alpha(a)] \quad \text{for all } a \in A.$$

Thus, the action of the composite mapping $\beta\alpha$ is “first α , then β ” (see the diagram on the next page), so the symbol $\beta\alpha$ must be read from right to left.⁵ Clearly, the composite $\alpha\beta$ cannot be formed unless $\beta(B) \subseteq A$. But even if $\alpha\beta$ and $\beta\alpha$ can both be defined, they need not be equal.

⁵Many authors write $\beta \circ \alpha$ for the composite mapping, but we use the simpler notation $\beta\alpha$.



Example 6. Let $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ and $\beta : \mathbb{R} \rightarrow \mathbb{R}$ be defined by $\alpha(x) = x + 1$ and $\beta(x) = x^2$ for all $x \in \mathbb{R}$. Find the action of $\beta\alpha$ and $\alpha\beta$ and conclude that $\alpha\beta \neq \beta\alpha$.

Solution. If $x \in \mathbb{R}$, then $\beta\alpha(x) = \beta[\alpha(x)] = \beta(x + 1) = (x + 1)^2$, whereas $\alpha\beta(x) = \alpha[\beta(x)] = \alpha(x^2) = x^2 + 1$. Clearly, $x \in \mathbb{R}$ exists with $\alpha\beta(x) \neq \beta\alpha(x)$, so $\alpha\beta \neq \beta\alpha$ by Theorem 1. \square

For a set A , the **identity map** $1_A : A \rightarrow A$ is defined by

$$1_A(a) = a \quad \text{for all } a \in A.$$

This mapping plays an important role; the notation 1_A is explained in part (1) of Theorem 3.

Theorem 3. Let $A \xrightarrow{\alpha} B \xrightarrow{\beta} C \xrightarrow{\gamma} D$ be mappings. Then

- (1) $\alpha 1_A = \alpha$ and $1_B \alpha = \alpha$.
- (2) $\gamma(\beta\alpha) = (\gamma\beta)\alpha$.
- (3) If α and β are both one-to-one (both onto), the same is true of $\beta\alpha$.

Proof. (1) If $a \in A$, then $\alpha 1_A(a) = \alpha[1_A(a)] = \alpha(a)$. Thus, $\alpha 1_A$ and α have the same action, that is, $\alpha 1_A = \alpha$. Similarly, $1_B \alpha = \alpha$.

(2) If $a \in A$: $[\gamma(\beta\alpha)](a) = \gamma[\beta\alpha(a)] = \gamma[\beta(\alpha(a))] = \gamma\beta[\alpha(a)] = [(\gamma\beta)\alpha](a)$.

(3) If α and β are one-to-one, suppose that $\beta\alpha(a) = \beta\alpha(a_1)$, where $a, a_1 \in A$. Thus, $\beta[\alpha(a)] = \beta[\alpha(a_1)]$, so $\alpha(a) = \alpha(a_1)$ because β is one-to-one. But then $a = a_1$ because α is one-to-one. This shows that $\beta\alpha$ is one-to-one.

Now assume that α and β are both onto. If $c \in C$, we have $c = \beta(b)$ for some $b \in B$ (because β is onto) and then $b = \alpha(a)$ for some $a \in A$ (because α is onto). Hence, $c = \beta[\alpha(a)] = \beta\alpha(a)$, proving that $\beta\alpha$ is onto. \blacksquare

We say that composition is **associative** because of the property $\gamma(\beta\alpha) = (\gamma\beta)\alpha$ in (2), and the composite is denoted simply as $\gamma\beta\alpha$. Note that the action of this mapping is

$$\gamma\beta\alpha(a) = \gamma[\beta[\alpha(a)]]$$

and so can be described as “first α , then β , then γ ” (see the proof of (2)).

Sometimes the action of one mapping *reverses* the action of another. For example, consider $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ and $\beta : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$\alpha(x) = 2x \quad \text{and} \quad \beta(x) = \frac{1}{2}x \quad \text{for all } x \in \mathbb{R}.$$

Then $\beta\alpha(x) = \beta[\alpha(x)] = \beta(2x) = \frac{1}{2}(2x) = x$ for all x ; that is, $\beta\alpha = 1_{\mathbb{R}}$. Hence, β *undoes* the action of α . Similarly, $\alpha\beta = 1_{\mathbb{R}}$. In this case, we say that α and β are inverses of each other.

In general, if $\alpha : A \rightarrow B$ is a mapping, a mapping $\beta : B \rightarrow A$ is called an **inverse** of α if

$$\beta\alpha = 1_A \quad \text{and} \quad \alpha\beta = 1_B.$$

Clearly, if β is an inverse of α , then automatically α is an inverse of β . As we show in Example 8, some mappings have no inverse. However, if β and β_1 are two inverses of α , we have $\beta_1\alpha = 1_A$ and $\alpha\beta = 1_B$. Hence,

$$\beta_1 = \beta_1 1_B = \beta_1(\alpha\beta) = (\beta_1\alpha)\beta = 1_A\beta = \beta$$

by Theorem 3, which proves Theorem 4.

Theorem 4. *If $\alpha : A \rightarrow B$ has an inverse, the inverse mapping is unique.*

A mapping $\alpha : A \rightarrow B$ that has an inverse is called an **invertible mapping**, and the inverse mapping is denoted α^{-1} . In this case, $\alpha^{-1} : B \rightarrow A$ is the unique mapping satisfying

$$\alpha^{-1}\alpha = 1_A \quad \text{and} \quad \alpha\alpha^{-1} = 1_B.$$

We can state these conditions as follows:

$$\alpha^{-1}[\alpha(a)] = a \quad \text{for all } a \in A \quad \text{and} \quad \alpha[\alpha^{-1}(b)] = b \quad \text{for all } b \in B.$$

These are the **Fundamental Identities** relating α and α^{-1} , and they show that the action of each of α and α^{-1} *undoes* the action of the other.

If we have $\alpha : A \rightarrow B$ and can somehow come up with a mapping $\beta : B \rightarrow A$ such that $\beta\alpha = 1_A$ and $\alpha\beta = 1_B$, then α is invertible and $\beta = \alpha^{-1}$. Here is an illustration.

Example 7. If $A = \{1, 2, 3\}$, define $\alpha : A \rightarrow A$ by $\alpha(1) = 2$, $\alpha(2) = 3$, and $\alpha(3) = 1$. Compute $\alpha^2 = \alpha\alpha$ and $\alpha^3 = \alpha\alpha\alpha$ and so find α^{-1} .

Solution. We have $\alpha^2(1) = 3$, $\alpha^2(2) = 1$, and $\alpha^2(3) = 2$, as the reader can verify, and so $\alpha^3(1) = 1$, $\alpha^3(2) = 2$, and $\alpha^3(3) = 3$. Thus, $\alpha^3 = 1_A$ and so $\alpha^2\alpha = 1_A = \alpha\alpha^2$. Hence, α is invertible and α^2 is the inverse; in symbols $\alpha^{-1} = \alpha^2$. \square

Theorem 5. *Let $\alpha : A \rightarrow B$ and $\beta : B \rightarrow C$ denote mappings.*

- (1) $1_A : A \rightarrow A$ is invertible and $1_A^{-1} = 1_A$.
- (2) If α is invertible, then α^{-1} is invertible and $(\alpha^{-1})^{-1} = \alpha$.
- (3) If α and β are both invertible, then $\beta\alpha$ is invertible and $(\beta\alpha)^{-1} = \alpha^{-1}\beta^{-1}$.

Proof. (1) This result follows because $1_A 1_A = 1_A$.

- (2) We have $\alpha^{-1}\alpha = 1_A$ and $\alpha\alpha^{-1} = 1_B$, so α is the inverse of α^{-1} .

(3) Compute $(\beta\alpha)(\alpha^{-1}\beta^{-1}) = \beta[\alpha\alpha^{-1}]\beta^{-1} = \beta 1_B \beta^{-1} = \beta\beta^{-1} = 1_C$. A similar calculation shows that $(\alpha^{-1}\beta^{-1})(\beta\alpha) = 1_A$, so $\alpha^{-1}\beta^{-1}$ is the inverse of $\beta\alpha$. Note the order of the factors. ■

Example 8. Define α and $\beta : \mathbb{N} \rightarrow \mathbb{N}$ by $\alpha(n) = n + 1$ for all $n \in \mathbb{N}$, and

$$\beta(n) = \begin{cases} 1, & \text{if } n = 0, \\ n - 1, & \text{if } n > 0. \end{cases}$$

Show that $\beta\alpha = 1_{\mathbb{N}}$ but that $\alpha\beta \neq 1_{\mathbb{N}}$. Conclude that α is not invertible.

Solution. We have $\beta\alpha(n) = \beta(n + 1) = (n + 1) - 1 = n$ for all $n \in \mathbb{N}$, so $\alpha\beta = 1_{\mathbb{N}}$. However, $\alpha\beta \neq 1_{\mathbb{N}}$ because, for example, $\alpha\beta(0) = \alpha(1) = 2$. Note that $0 \notin \alpha(\mathbb{N})$, so α is not onto. Hence, α is not invertible by Theorem 6. □

Theorem 6. Invertibility Theorem. A mapping $\alpha : A \rightarrow B$ is invertible if and only if it is both one-to-one and onto (that is, α is a bijection).

Proof. Assume α^{-1} exists. If $\alpha(a) = \alpha(a_1)$, then $a = \alpha^{-1}[\alpha(a)] = \alpha^{-1}[\alpha(a_1)] = a_1$ by one of the fundamental identities. Hence, α is one-to-one. If $b \in B$, then $b = \alpha[\alpha^{-1}(b)]$ by the other fundamental identity, and so α is onto.

Conversely, assume α is onto and one-to-one. Given $b \in B$, there exists $a \in A$ such that $\alpha(a) = b$ (because α is onto) and a is unique (because α is one-to-one, verify). Hence, we may define $\beta : B \rightarrow A$ by $\beta(b) = a$, where a is the unique element of A with $\alpha(a) = b$. Thus, $\alpha\beta(b) = \alpha(a) = b$ for each $b \in B$, so $\alpha\beta = 1_B$. If $a \in A$, write $\alpha(a) = b$. Hence, $\beta(b) = a$ by the definition of β , so $\beta\alpha(a) = \beta(b) = a$. This means that $\beta\alpha = 1_A$, so β is the inverse of α . ■

Theorem 6 is important because it can show that a mapping is invertible even though no simple formula for the inverse is known. For example, we can show (using calculus) that the function $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ given by $\alpha(x) = x^3 + 2x$ is one-to-one and onto. But a simple formula for α^{-1} is not easy to write.

Exercises 0.3

1. In each case, determine whether α is a well-defined mapping. Justify your answer.

- (a) $\alpha : \mathbb{N} \rightarrow \mathbb{N}$ defined by $\alpha(n) = -n$ for all $n \in \mathbb{N}$.
- (b) $\alpha : \mathbb{N} \rightarrow \mathbb{N}$ defined by $\alpha(n) = 1$ for all $n \in \mathbb{N}$.
- (c) $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\alpha(x) = \sqrt{x}$ for all $x \in \mathbb{R}$.
- (d) $\alpha : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ defined by $\alpha(x, y) = x + y$ for all $(x, y) \in \mathbb{R} \times \mathbb{R}$.
- (e) $\alpha : \mathbb{R} \rightarrow \mathbb{R} \times \mathbb{R}$ defined by $\alpha(xy) = (x, y)$ for all $xy \in \mathbb{R}$.

- (f) $\alpha : \{1, 2, 3\} \rightarrow \{a, b, c\}$ defined by the diagram
- | | | |
|---|-------------------|---|
| 1 | \longrightarrow | a |
| 2 | \nearrow | b |
| 3 | \searrow | c |
-
- (g) $\alpha : \{1, 2, 3\} \rightarrow \{a, b, c\}$ defined by the diagram
- | | | |
|---|-------------------|---|
| 1 | \longrightarrow | a |
| 2 | \nearrow | b |
| 3 | \searrow | c |

2. In each case, state whether the mapping is onto, one-to-one, or bijective. Justify your answer.

- (a) $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\alpha(x) = 3 - 4x$.
 (b) $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\alpha(x) = 1 + x^2$.
 (c) $\alpha : \mathbb{N} \rightarrow \mathbb{N}$ defined by

$$\alpha(n) = \begin{cases} \frac{n+1}{2}, & \text{if } n \text{ is odd,} \\ \frac{n}{2}, & \text{if } n \text{ is even.} \end{cases}$$

 (d) $\alpha : \mathbb{Z} \times \mathbb{Z}^+ \rightarrow \mathbb{Q}$ defined by $\alpha(n, m) = \frac{n}{m}$.
 (e) $\alpha : \mathbb{R} \rightarrow \mathbb{R} \times \mathbb{R}$ defined by $\alpha(x) = (x+1, x-1)$.
 (f) $\alpha : A \times B \rightarrow A$ defined by $\alpha(a, b) = a$. (Assume that $A \neq \emptyset \neq B$.)
 (g) $\alpha : A \rightarrow A \times B$ defined by $\alpha(a) = (a, b_0)$, where $b_0 \in B$ is fixed and $A \neq \emptyset$.
3. Let $A \xrightarrow{\alpha} B \xrightarrow{\beta} C$ be mappings.
- (a) If $\beta\alpha$ is onto, show that β is onto.
 - (b) If $\beta\alpha$ is one-to-one, show that α is one-to-one.
 - (c) If $\beta\alpha$ is one-to-one and α is onto, show that β is one-to-one.
 - (d) If $\beta\alpha$ is onto and β is one-to-one, show that α is onto.
 - (e) If $\beta_1 : B \rightarrow C$ satisfies $\beta\alpha = \beta_1\alpha$ and α is onto, show that $\beta = \beta_1$.
 - (f) If $\alpha_1 : A \rightarrow B$ satisfies $\beta\alpha = \beta\alpha_1$ and β is one-to-one, show that $\alpha = \alpha_1$.
4. For $\alpha : A \rightarrow A$, show that $\alpha^2 = 1_A$ if and only if α is invertible and $\alpha^{-1} = \alpha$.
5. (a) For $A \xrightarrow{\alpha} A$, show that $\alpha^2 = \alpha$ if and only if $\alpha(x) = x$ for all $x \in \alpha(A)$.
 (b) If $A \xrightarrow{\alpha} A$ satisfies $\alpha^2 = \alpha$, show that α is onto if and only if α is one-to-one. Describe α in this case.
 (c) Let $A \xrightarrow{\beta} B \xrightarrow{\gamma} A$ satisfy $\gamma\beta = 1_A$. If $\alpha = \beta\gamma$, show that $\alpha^2 = \alpha$.
6. If $|A| \geq 2$ and $\alpha : A \rightarrow A$ satisfies $\alpha\beta = \beta\alpha$ for all $\beta : A \rightarrow A$, prove that $\alpha = 1_A$.
7. In each case, verify that α^{-1} exists and describe its action.
- (a) $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ defined by $\alpha(x) = ax + b$, where $0 \neq a \in \mathbb{R}$ and $b \in \mathbb{R}$.
 - (b) $\alpha : \mathbb{R} \rightarrow \{x \in \mathbb{R} \mid x > 1\}$ defined by $\alpha(x) = 1 + x^2$.
 - (c) $\alpha : \mathbb{N} \rightarrow \mathbb{N}$ defined by $\alpha(n) = \begin{cases} n+1, & \text{if } n \text{ is even,} \\ n-1, & \text{if } n \text{ is odd.} \end{cases}$
 - (d) $\alpha : A \times B \rightarrow B \times A$ defined by $\alpha(a, b) = (b, a)$.
8. Let $A \xrightarrow{\alpha} B \xrightarrow{\beta} A$ satisfy $\beta\alpha = 1_A$. If either α is onto or β is one-to-one, show that each of them is invertible and that each of them is the inverse of the other.
9. Let $A \xrightarrow{\alpha} B \xrightarrow{\beta} A$ satisfy $\beta\alpha = 1_A$. If A and B are finite sets with $|A| = |B|$, show that $\alpha\beta = 1_B$, $\alpha = \beta^{-1}$, and $\beta = \alpha^{-1}$. (Compare your answer with the solution of Example 8.)
10. For $A \xrightarrow{\alpha} B \xrightarrow{\beta} A$, show that both $\alpha\beta$ and $\beta\alpha$ have inverses if and only if both α and β have inverses.
11. Let M denote the set of all mappings $\alpha : \{1, 2\} \rightarrow B$. Define $\varphi : M \rightarrow B \times B$ by $\varphi(\alpha) = (\alpha(1), \alpha(2))$. Show that φ is a bijection and find the action of φ^{-1} .
12. A mapping $\delta : A \rightarrow B$ is called a **constant map** if there exists $b_0 \in B$ such that $\delta(a) = b_0$ for all $a \in A$. Show that a mapping $\delta : A \rightarrow B$ is constant if and only if $\delta\alpha = \delta$ for all $\alpha : A \rightarrow A$.
13. If $|A| = n$ and $|B| = m$, show that there are m^n mappings $A \rightarrow B$.
14. Show that the following conditions are equivalent for a mapping $\alpha : A \rightarrow B$, where A and B are nonempty.
- (a) α is one-to-one.
 - (b) There exists $\beta : B \rightarrow A$ such that $\beta\alpha = 1_A$.
 - (c) If $\gamma : C \rightarrow A$ and $\delta : C \rightarrow A$ satisfy $\alpha\gamma = \alpha\delta$, then $\gamma = \delta$.

15. Show that the following conditions are equivalent for a mapping $\alpha : A \rightarrow B$, where A and B are nonempty.
- α is onto.
 - There exists $\beta : B \rightarrow A$ such that $\alpha\beta = 1_B$.
 - If $\gamma : B \rightarrow C$ and $\delta : B \rightarrow C$ satisfy $\gamma\alpha = \delta\alpha$, then $\gamma = \delta$.
16. If $A \neq \emptyset$ and $P(A) = \{X \mid X \subseteq A\}$, show there is no onto mapping $\alpha : A \rightarrow P(A)$.
[Hint. Let $R = \{r \in A \mid r \notin \alpha(r)\}$. If $R = \alpha(a)$, is $a \in R$ or $a \notin R$?]

0.4 EQUIVALENCES

It often happens that elements of a set are alike in some respect, but they are not necessarily equal. For example, similar triangles are alike in that they have the same angles, but they need not be equal in size. For another example, two subsets of a finite set may be regarded as alike if they have the same number of elements. The concept of an equivalence relation unifies such examples in a useful way.

If A is a set, a subset \equiv of $A \times A$ is called a **relation** on A . For elements a and b in A , we customarily write

$$a \equiv b \quad \text{to mean} \quad (a, b) \text{ is an element of the set } \equiv$$

and we write $a \not\equiv b$ when (a, b) is not in \equiv .

A relation \equiv on a set A is called an **equivalence** on A if it satisfies the following conditions, where a , b , and c denote elements of A :

- (1) $a \equiv a$ for all $a \in A$ (reflexive property),
- (2) If $a \equiv b$, then $b \equiv a$ (symmetric property),
- (3) If $a \equiv b$ and $b \equiv c$, then $a \equiv c$ (transitive property).

If \equiv is an equivalence on a set A , the statement $a \equiv b$ is read as “ a is equivalent to b ”. Certainly, equality is an example of an equivalence, and the notation \equiv reflects the idea that an equivalence relation is a weakened form of equality. Intuitively, $a \equiv b$ holds when a and b are *alike* in some sense. Thus, given an element a of A , the set of all elements equivalent to a plays a central role in revealing the structure of the equivalence relation.

More formally, let \equiv be an equivalence on a set A . Given $a \in A$, the **equivalence class** $[a]$ of a is defined as the set of all elements of A that are equivalent to a :

$$[a] = \{x \in A \mid x \equiv a\}.$$

The equivalence class $[a]$ is said to be **generated** by a .

Examples 1–5 illustrate equivalences. In most cases, we leave verification of the three defining properties to the reader.

Example 1. Equality is an equivalence on any set A . If $a \in A$, the equivalence class of a is $[a] = \{x \in A \mid x = a\} = \{a\}$, the singleton.

Example 2. Being parallel is an equivalence on the set of lines in the plane. The equivalence class of a given line consists of all lines parallel to it.

Example 3. If X and Y are subsets of a finite set U , write $X \equiv Y$ to mean $|X| = |Y|$. Then \equiv is an equivalence on the set of subsets of U , and $[X]$ consists of all subsets with the same number of elements as X .

Example 4. Let $\alpha : A \rightarrow B$ be a mapping. If a and a_1 are elements of A , write $a \equiv a_1$ to mean $\alpha(a) = \alpha(a_1)$. Then \equiv is an equivalence on A , called the **kernel** equivalence of α , and $[a] = \{x \in A \mid \alpha(x) = \alpha(a)\}$ for each $a \in A$.

Example 5. If m and n are integers, define $m \equiv n$ to mean that $m - n$ is even. Then \equiv is an equivalence on \mathbb{Z} . (Proof of transitivity: If $m \equiv n$ and $n \equiv k$, then both $m - n$ and $n - k$ are even, so $m - k = (m - n) + (n - k)$ is also even. Thus, $m \equiv k$.) In this case,

- $[0] = \{x \in \mathbb{Z} \mid x \equiv 0\}$ is the set of even integers, and
- $[1] = \{x \in \mathbb{Z} \mid x \equiv 1\}$ is the set of odd integers.

Moreover, it is not difficult to verify that $[m] = [0]$ if m is even and $[m] = [1]$ if m is odd, and so $[0]$ and $[1]$ are the *only* equivalence classes. \square

We describe equivalences as the one in Example 5 in more detail in Section 1.3.

Theorem 1 collects the basic properties of equivalence classes.

Theorem 1. Let \equiv be an equivalence on a set A and let a and b denote elements of A . Then

- (1) $a \in [a]$ for every $a \in A$.
- (2) $[a] = [b]$ if and only if $a \equiv b$.
- (3) If $a \in [b]$ then $[a] = [b]$.
- (4) If $[a] \neq [b]$ then $[a] \cap [b] = \emptyset$.

Proof. (1) This is clear because $a \equiv a$ for all $a \in A$ by the reflexive property.

(2) If $[a] = [b]$, then $a \in [b]$ by (1), so $a \equiv b$. Conversely, assume $a \equiv b$. If $x \in [a]$, then $x \equiv a$, so, since $a \equiv b$, we have $x \equiv b$ by transitivity. Thus, $x \in [b]$ and we have proved that $[a] \subseteq [b]$. The other inclusion $[b] \subseteq [a]$ follows in the same way because $b \equiv a$ by symmetry. Hence $[a] = [b]$.

(3) If $a \in [b]$, then $a \equiv b$, so $[a] = [b]$ by (2).

(4) We argue by contradiction. If $[a] \neq [b]$, we assume on the contrary that $[a] \cap [b] \neq \emptyset$, say $x \in [a] \cap [b]$. Then $x \equiv a$ and $x \equiv b$, so $a \equiv b$ by the symmetric and transitive properties. But then $[a] = [b]$ by (2), a contradiction. \blacksquare

The view that an equivalence is a weakened version of equality is upheld by (2) of Theorem 1. However, the equality is for equivalence classes rather than elements. Property (2) is used several times in this book.

Partitions

Theorem 1 leads to a useful description of equivalence relations. Two sets X and Y are called **disjoint** if they have no element in common (that is, $X \cap Y = \emptyset$), and a family of sets is called **pairwise disjoint** if any two (distinct) sets in the family are disjoint.

If A is a nonempty set, a family \mathcal{P} of subsets of A is called a **partition** of A (and the sets in \mathcal{P} are called the **cells** of the partition) if

- (1) No cell is empty.
- (2) The cells are pairwise disjoint.
- (3) Every element of A belongs to some cell.

If \mathcal{P} is a partition of A , (2) and (3) clearly imply that each element of A lies in *exactly one* cell of \mathcal{P} .

The simplest partition of A is the **trivial partition** $\mathcal{P} = \{A\}$ with just one cell: A itself. At the other extreme is the **singleton partition** $\mathcal{P} = \{\{a\} \mid a \in A\}$, where every cell is a singleton.

Example 6. The set $A = \{1, 2, 3\}$ has five partitions:

$$\{A\} \quad \{\{1, 2\}, \{3\}\} \quad \{\{1, 3\}, \{2\}\} \quad \{\{2, 3\}, \{1\}\} \quad \{\{1\}, \{2\}, \{3\}\}$$

Partitions of a set A give rise to equivalences on A in a natural way. If \mathcal{P} is a partition of the nonempty set A , and if a and b are elements of A , we define $a \equiv b$ to mean that a and b are in the same cell of \mathcal{P} . Then \equiv is reflexive because each $a \in A$ lies in *some* cell, so $a \equiv a$. The relation \equiv is obviously symmetric. To show that it is transitive, we let $a \equiv b$ and $b \equiv c$. Because b lies in a *unique* cell, a and c are in that same cell; that is, $a \equiv c$. Hence, \equiv is an equivalence on A , and we say that it is the equivalence **afforded** by the partition \mathcal{P} . Surprisingly, *every* equivalence on A arises in this way.

Theorem 2. Partition Theorem. *If \equiv is any equivalence on a nonempty set A , the family of all equivalence classes is a partition of A that affords \equiv .*

Proof. The equivalence classes are nonempty and pairwise disjoint by (1) and (4) of Theorem 1, and every element of A belongs to some class (the one it generates). Hence, the equivalence classes are the cells of a partition. To show that this partition affords \equiv , it is enough to show that two elements a and b are equivalent if and only if they belong to the same equivalence class. If $a \equiv b$, then $[a] = [b]$ by (2) of Theorem 1, so a and b belong to this common class. Conversely, if a and b belong to class $[c]$, then $[a] = [c] = [b]$ by (3) of Theorem 1, so $a \equiv b$. ■

Theorem 2 shows that partitions of A and equivalences on A are actually two ways of looking at the same phenomenon—classifying the elements of A . On the one hand, we classify them by declaring which pairs of elements are equivalent; on the other hand, we classify them by partitioning A into disjoint cells.

For example, equality on a set A is the equivalence afforded by the singleton partition of A . At the other extreme, the trivial partition $\{A\}$ affords the equivalence that declares that *any* two elements of A are equivalent.

If \equiv is an equivalence on A , the set of all equivalence classes is called the **quotient set** and denoted A_{\equiv} . Hence,

$$A_{\equiv} = \{[a] \mid a \in A\}.$$

The mapping

$$\varphi : A \rightarrow A_{\equiv} \quad \text{given by} \quad \varphi(a) = [a] \quad \text{for all } a \in A$$

is called the **natural mapping**. The natural mapping φ is clearly onto and (3) of Theorem 1 shows that $\varphi(a) = \varphi(a_1)$ if and only if $a \equiv a_1$. In other words, \equiv is the

kernel equivalence of φ (see Example 4). This proves the following consequence of the partition theorem.

Corollary. *Every equivalence on a set A is the kernel equivalence of some onto mapping with A as domain.*

The fact that the same equivalence class can have different generators leads to a minor difficulty when we are defining a mapping whose domain is a quotient set. This problem usually arises in the following way. Suppose that \equiv is an equivalence on a set A and that a mapping

$$\alpha : A \rightarrow B$$

is given. If we write $A_{\equiv} = \{[a] \mid a \in A\}$ as before, we are often interested in defining

$$\sigma : A_{\equiv} \rightarrow B \quad \text{by} \quad \sigma([a]) = \alpha(a)$$

for each equivalence class $[a]$ in A_{\equiv} . The question is whether σ is a *mapping*. The problem is that a given equivalence class C could be generated by distinct elements of A :

$$C = [a] = [a_1],$$

where $a \neq a_1$. Then $\sigma(C)$ will be $\alpha(a)$ or $\alpha(a_1)$, depending on whether we use $C = [a]$ or $C = [a_1]$. Clearly, if the action of σ is to make sense

$$[a] = [a_1] \quad \text{must imply that} \quad \alpha(a) = \alpha(a_1).$$

Then the assignment of $\sigma([a]) = \alpha(a)$ does not depend on which element a generates the equivalence class. We express this conclusion by saying that σ is **well defined** by this formula.

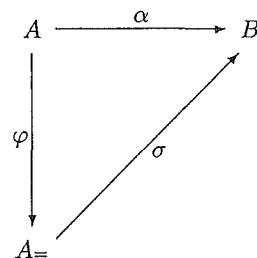
Example 7. Let \equiv be the equivalence on \mathbb{Z} defined by $m \equiv n$ if $m - n$ is even (Example 5). Show that the mapping $\sigma : \mathbb{Z}_{\equiv} \rightarrow \{1, -1\}$ is well defined by $\sigma([n]) = (-1)^n$. Then show that σ is a bijection.

Solution. To show that σ is well defined, we must show that $[m] = [n]$ implies $(-1)^m = (-1)^n$. But $[m] = [n]$ implies $m \equiv n$ by (2) of Theorem 1 so $m - n$ is even. Hence, both m and n are even or both are odd, and $(-1)^m = (-1)^n$ follows. Thus, σ is well defined. Verification that σ is one-to-one is the converse of the argument that it is well-defined: If $\sigma([m]) = \sigma([n])$, then $(-1)^m = (-1)^n$, so m and n are both even or both odd. Either way $m - n$ is even, so $m \equiv n$. This means that $[m] = [n]$ by Theorem 1, proving that σ is one-to-one. As σ is clearly onto, it is a bijection. Note that this result shows that $|\mathbb{Z}_{\equiv}| = 2$, a fact confirmed in a different way in Example 5. \square

Exercises 0.4

- In each case, decide whether the relation \equiv is an equivalence on A . Give reasons for your answer. If it is an equivalence, describe the equivalence classes.
 - $A = \{-2, -1, 0, 1, 2\}$; $a \equiv b$ means that $a^3 - a = b^3 - b$.
 - $A = \{-1, 0, 1\}$; $a \equiv b$ means that $a^2 = b^2$.
 - $A = \{x \in \mathbb{R} \mid x > 0\}$; $x \equiv y$ means that $xy = 1$.

- (d) $A = \mathbb{N}$; $a \equiv b$ means that $a \leq b$.
 (e) $A = \mathbb{N}$; $a \equiv b$ means that $b = ka$ for some integer k .
 (f) $A =$ the set of all subsets of $\{1, 2, 3\}$; $X \equiv Y$ means that $|X| = |Y|$.
 (g) $A =$ the set of lines in the plane; $x \equiv y$ means x is perpendicular to y .
 (h) $A = \mathbb{R} \times \mathbb{R}$; $(x, y) \equiv (x_1, y_1)$ means that $x^2 + y^2 = x_1^2 + y_1^2$.
 (i) $A = \mathbb{R} \times \mathbb{R}$; $(x, y) \equiv (x_1, y_1)$ means that $y - 3x = y_1 - 3x_1$.
2. Let $U = \{1, 2, 3\}$ and $A = U \times U$. In each case, show that \equiv is an equivalence on A and find the quotient set A_{\equiv} .
- (a) $(a, b) \equiv (a_1, b_1)$ if $a + b = a_1 + b_1$.
 - (b) $(a, b) \equiv (a_1, b_1)$ if $ab = a_1 b_1$.
 - (c) $(a, b) \equiv (a_1, b_1)$ if $a = a_1$.
 - (d) $(a, b) \equiv (a_1, b_1)$ if $a - b = a_1 - b_1$.
3. In each case, show that \equiv is an equivalence on A and find a (well-defined) bijection $\sigma : A_{\equiv} \rightarrow B$.
- (a) $A = \mathbb{Z}$; $m \equiv n$ means that $m^2 = n^2$; $B = \mathbb{N}$.
 - (b) $A = \mathbb{R} \times \mathbb{R}$; $(x, y) \equiv (x_1, y_1)$ means that $x^2 + y^2 = x_1^2 + y_1^2$; $B = \{x \in \mathbb{R} \mid x \geq 0\}$.
 - (c) $A = \mathbb{R} \times \mathbb{R}$; $(x, y) \equiv (x_1, y_1)$ means that $y = y_1$; $B = \mathbb{R}$.
 - (d) $A = \mathbb{R}^+ \times \mathbb{R}^+$; $(x, y) \equiv (x_1, y_1)$ means that $y/x = y_1/x_1$; $B = \{x \in \mathbb{R} \mid x > 0\}$.
 - (e) $A = \mathbb{R}$; $x \equiv y$ means that $x - y \in \mathbb{Z}$; $B = \{x \in \mathbb{R} \mid 0 \leq x < 1\}$.
 - (f) $A = \mathbb{Z}$; $m \equiv n$ means that $m^2 - n^2$ is even; $B = \{0, 1\}$.
4. Find all partitions of $A = \{1, 2, 3, 4\}$.
5. Let $\mathcal{P}_1 = \{C_1, C_2, \dots, C_m\}$ and $\mathcal{P}_2 = \{D_1, D_2, \dots, D_n\}$ be partitions of a set A .
- (a) Show that $\mathcal{P} = \{C_i \cap D_j \mid C_i \cap D_j \neq \emptyset\}$ is also a partition of A .
 - (b) If \equiv_1, \equiv_2 , and \equiv denote the equivalences afforded by $\mathcal{P}_1, \mathcal{P}_2$, and \mathcal{P} , respectively, describe \equiv in terms of \equiv_1 and \equiv_2 .
6. Let \equiv and \sim be two equivalences on the same set A .
- (a) If $a \equiv a_1$ implies that $a \sim a_1$, show that each \sim equivalence class is partitioned by the \equiv equivalence classes it contains.
 - (b) Define \cong on A by writing $a \cong a_1$ if and only if both $a \equiv a_1$ and $a \sim a_1$. Show that \cong is an equivalence and describe the \cong equivalence classes in terms of the \equiv and \sim equivalence classes.
7. In each case, determine whether $\alpha : \mathbb{Q}^+ \rightarrow \mathbb{Q}$ is well defined, where \mathbb{Q}^+ is the set of positive rational numbers. Support your answer.
- (a) $\alpha(\frac{n}{m}) = n$ (b) $\alpha(\frac{n}{m}) = \frac{n-m}{n+m}$ (c) $\alpha(\frac{n}{m}) = m+n$ (d) $\alpha(\frac{n}{m}) = \frac{5m+7n}{3n+m}$
8. Define \equiv and \sim on \mathbb{R} by $x \equiv y$ if $x - y \in \mathbb{Z}$ and by $x \sim y$ if $x - y \in \mathbb{Q}$.
- (a) Show that \equiv and \sim are equivalences.
 - (b) Show that $\alpha : \mathbb{R}_{\equiv} \rightarrow \mathbb{R}_{\sim}$ is well defined and onto if $\alpha([x]_{\equiv}) = [x]_{\sim}$. Is the mapping α one-to-one?
9. For a mapping $\alpha : A \rightarrow B$, let \equiv denote the kernel equivalence of α and let $\varphi : A \rightarrow A_{\equiv}$ denote the natural mapping. Define
- $$\sigma : A_{\equiv} \rightarrow B \quad \text{by} \quad \sigma([a]) = \alpha(a)$$
- for all equivalence classes $[a]$ in A_{\equiv} .
- (a) Show that σ is well defined and one-to-one, onto if α is onto.



- (b) Show that $\alpha = \sigma\varphi$, so that α is the composite of an onto mapping followed by a one-to-one mapping.
- (c) If $\alpha(A)$ is a finite set, show that the set A_{\equiv} of equivalence classes is also finite and that $|A_{\equiv}| = |\alpha(A)|$. (This result is called the **Bijection Theorem**).
- (d) In each case, find $|A_{\equiv}|$ for the given mapping α .
- $A = U \times U$ with $U = \{1, 2, 3, 4, 6, 12\}$, $\alpha : A \rightarrow \mathbb{Q}$ defined by $\alpha(n, m) = n/m$.
 - $A = \{n \in \mathbb{Z} \mid 1 \leq n \leq 99\}$, $\alpha : A \rightarrow \mathbb{N}$ defined by $\alpha(n) =$ the sum of the digits of n .
10. Let $A = \{\alpha \mid \alpha : P \rightarrow Q$ is a mapping}. Given $p \in P$, define \equiv on A by $\alpha \equiv \beta$ if $\alpha(p) = \beta(p)$.
- Show that \equiv is an equivalence on A .
 - Find a mapping $\lambda : A \rightarrow Q$ such that \equiv is the kernel equivalence of λ .
 - If $|Q| = n$, how many equivalence classes does \equiv have? [Hint: Exercise 9.]

Chapter 1

Integers and Permutations

God made the integers, and all the rest is the work of man.

—Leopold Kronecker

The use of arithmetic is a basic aspect of human culture. Anthropologists tell us that even the most primitive societies, because of their desire to count objects, have developed some sort of terminology for the numbers 1, 2, and 3, although many go no further. As a culture develops, it needs more sophisticated counting to deal with commerce, warfare, the calendar, and so on. This leads to methods of recording numbers often (but by no means always) based on groups of 10, presumably from counting on the fingers. Then the recording of numbers by making marks or notches becomes important (in bookkeeping, for example), and a variety of systems have been constructed for doing so. Many of these systems were not very useful for adding or multiplying (try multiplying with Roman numerals), and the development of our positional system, originating with the Babylonians using base 60 rather than 10, was a great advance.

In this chapter we assume the validity of the elementary arithmetic properties of the integers and use them to derive some more subtle facts related to divisibility and primes. Then two fundamental algebraic systems are described: the integers modulo n and the permutations of the set $\{1, 2, \dots, n\}$. These are, respectively, excellent examples of *rings* and *groups*, two of the basic algebraic structures presented in detail in Chapters 2 and 3.

1.1 INDUCTION

Great fleas have little fleas upon their backs to bite 'em, And little fleas have lesser fleas,
and so ad infinitum.

—Augustus De Morgan

Consider the sequence of equations:

$$\begin{aligned} 1 &= 1 \\ 1 + 3 &= 4 \\ 1 + 3 + 5 &= 9 \\ 1 + 3 + 5 + 7 &= 16 \\ &\vdots \end{aligned}$$

It is clear there is a pattern. The right sides are the squares $1^2, 2^2, 3^2, 4^2, \dots$, and, when the right side is n^2 , the left side is the sum of the first n odd integers. As the n th odd integer is $2n - 1$, the following expression is true for $n = 1, 2, 3$, and 4:

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2. \quad (p_n)$$

Now it is almost irresistible to ask whether the statement (p_n) is true for *every* $n \geq 1$. There is no hope of separately verifying all these statements, because there are infinitely many of them. A more subtle approach is required.

The idea is to prove that $p_k \Rightarrow p_{k+1}$ for every $k \geq 1$. Then the fact that p_1 is true implies that p_2 is true, which in turn implies that p_3 is true, then p_4 , and so on. This is one of the most important axioms for the integers.

Principle of Mathematical Induction⁶. Let p_n be a statement for each integer $n \geq 1$. Suppose that the following conditions are satisfied:

- (1) p_1 is true.
- (2) $p_k \Rightarrow p_{k+1}$ for every $k \geq 1$.

Then p_n is true for every $n \geq 1$.

In the proof that $p_k \Rightarrow p_{k+1}$, we assume that p_k is true and use it to prove that p_{k+1} is also true. The assumption that p_k is true is called the **induction hypothesis**.

For a graphic illustration, consider an infinite row of dominoes labeled 1, 2, 3, ... standing so that if one is knocked over, it will knock the next one over. If p_k is the statement that domino k falls over, this means that $p_k \Rightarrow p_{k+1}$ for each $k \geq 1$. The principle of induction asserts that knocking domino 1 over causes them all to fall.

As another illustration, let p_n be the statement $1 + 3 + 5 + \cdots + (2n - 1) = n^2$ mentioned above. Then p_1 has already been verified. To prove that $p_k \Rightarrow p_{k+1}$ for each $k \geq 1$, we assume that p_k is true (the induction hypothesis) and use it to simplify the left side of the sum p_{k+1} :

$$1 + 3 + 5 + \cdots + (2k - 1) + (2k + 1) = k^2 + (2k + 1) = (k + 1)^2.$$

⁶One of the earliest uses of the principle is in the work of Francesco Maurolico in the 16th century. Augustus De Morgan coined the name *mathematical induction* in 1838.

This expression shows that p_{k+1} is true and hence, by the induction principle, that p_n is true for all $n \geq 1$.

Example 1. Prove **Gauss' Formula**⁷: $1 + 2 + \cdots + n = \frac{1}{2}n(n + 1)$ for all $n \geq 1$.

Solution. Let p_n denote the statement $1 + 2 + \cdots + n = \frac{1}{2}n(n + 1)$. Then p_1 is true because $1 = \frac{1}{2}(1 + 1)$. If we assume that p_k is true for some $k \geq 1$, we get

$$1 + 2 + 3 + \cdots + k + (k + 1) = \frac{1}{2}k(k + 1) + (k + 1) = \frac{1}{2}(k + 1)(k + 2),$$

which shows that p_{k+1} is true. Hence, p_n is true for all $n \geq 1$ by the principle of mathematical induction. \square

Example 2 gives an inductive proof of a useful formula for the sum of a geometric series $1 + x + \cdots + x^n$. We use the convention that $x^0 = 1$ for all numbers x .

Example 2. If x is any real number, show that

$$(1 - x)(1 + x + \cdots + x^{n-1}) = 1 - x^n, \quad \text{for all } n \geq 1.$$

Solution. Let p_n be the given statement. Then p_1 is $(1 - x)1 = 1 - x^1$, which is true. If we assume that p_k is true for some $k \geq 1$, then the left side of p_{k+1} becomes

$$\begin{aligned} (1 - x)(1 + x + \cdots + x^{k-1} + x^k) &= (1 - x)(1 + x + \cdots + x^{k-1}) + (1 - x)x^k \\ &= (1 - x^k) + (1 - x)x^k \\ &= 1 - x^{k+1}. \end{aligned}$$

This proves that p_{k+1} is true and so completes the induction. \square

Example 3. Let w_n denote the number of n -letter words that can be formed using only the letters a and b . Show that $w_n = 2^n$ for all $n \geq 1$.

Solution. Clearly, a and b are the only such words with one letter, so $w_1 = 2 = 2^1$. If $k \geq 1$, we obtain each such word of $k + 1$ letters by adjoining an a or a b to a word of k letters, and there are w_k of each type. Hence, $w_{k+1} = 2w_k$ for each $k \geq 1$ so, if we assume inductively that $w_k = 2^k$, we get $w_{k+1} = 2w_k = 2 \cdot 2^k = 2^{k+1}$, as required. \square

The principle of induction starts at 1 in the sense that if p_1 is true and $p_k \Rightarrow p_{k+1}$ for all $k \geq 1$, then p_k is true for all $k \geq 1$. There is nothing special about 1.

Theorem 1. If m is any integer, let $p_m, p_{m+1}, p_{m+2}, \dots$ be statements such that

- (1) p_m is true.
- (2) $p_k \Rightarrow p_{k+1}$ for every $k \geq m$.

Then p_n is true for each $n \geq m$.

⁷This formula was probably known to the ancient Greeks. However, the great mathematician Carl Friedrich Gauss is said to have derived a special case of the formula ($n = 100$) at age 7 by writing the sum $1 + 2 + \cdots + 100$ in two parts:

$$\begin{array}{r} 1 + 2 + \cdots + 49 + 50 \\ 100 + 99 + \cdots + 52 + 51 \end{array}$$

and observing that each pair of terms, $1 + 100, 2 + 99, \dots, 50 + 51$, adds to 101. As there are 50 such pairs, the sum is $50 \cdot 101 = 5050$.

Proof. Let $t_n = p_{m+n-1}$ for each $n \geq 1$. Then $t_1 = p_m$ is true, and $t_k \Rightarrow t_{k+1}$ because $p_{m+k-1} \Rightarrow p_{m+k}$. Hence, t_n is true for all $n \geq 1$ by induction; that is, p_n is true for all $n \geq m$. \blacksquare

Example 4. If $n \geq 8$, show that any postage of n cents can be made exactly using only 3- and 5 cent stamps.

Solution. The assertion clearly holds if $n = 8$. If it holds for some $k \geq 8$, we consider two cases:

Case 1. One or more 5 cent stamps are used to make up k cents postage.

Then replace one of them with two 3 cent stamps.

Case 2. Three or more 3 cent stamps are used to make up k cents postage.

Then replace three of them with two 5 cent stamps.

Because one of these cases must occur (as $k \geq 8$), the assertion holds for $k + 1$ cents in both cases and the induction goes through. \square

If $n \geq 1$ is an integer, the integer $n!$ (read n -factorial) is defined to be the product

$$n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$$

of all the integers from n to 1. Thus, $1! = 1$, $2! = 2$, $3! = 6$, and so on. Clearly,

$$(n+1)! = (n+1)n!, \quad \text{for each } n \geq 1,$$

which we extend to $n = 0$ by defining

$$0! = 1.$$

Example 5. Show that $2^n < n!$ for all $n \geq 4$.

Solution. If p_k is the statement $2^k < k!$, note that p_1, p_2 , and p_3 are actually false, but p_4 is true because $2^4 = 16 < 24 = 4!$. If p_k is true where $k \geq 4$, then $2^k < k!$ so

$$2^{k+1} = 2 \cdot 2^k < 2 \cdot k! < (k+1)k! = (k+1)!$$

Hence, p_{k+1} is true and the induction is complete. \square

Let n and r be integers with $0 < r \leq n$. The **binomial coefficient** $\binom{n}{r}$ is defined as follows:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

As $0! = 1$, we have $\binom{n}{0} = 1 = \binom{n}{n}$ and $\binom{n}{2} = \frac{n(n-1)}{2}$. It is easy to verify that

$$\binom{n}{r} = \binom{n}{n-r}, \quad \text{whenever } 0 \leq r \leq n.$$

We leave the proof of the following formula (the **Pascal identity**) as Exercise 13.

$$\binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}, \quad \text{whenever } 1 \leq r \leq n.$$

The name honors Blaise Pascal. The identity leads to a way of displaying the binomial coefficients known as **Pascal's triangle**:

$$\begin{array}{c}
 & & & 1 \\
 & & 1 & 1 \\
 & 1 & 2 & 1 \\
 1 & 3 & 3 & 1 \\
 1 & 4 & 6 & 4 & 1 \\
 & & \vdots
 \end{array}$$

The n^{th} row of the triangle is $\binom{n}{0} \binom{n}{1} \binom{n}{2} \cdots \binom{n}{n-1} \binom{n}{n}$, starting at $n = 0$. The Pascal identity shows that each entry in a given row (except at the ends) can be found by adding the two entries adjacent to it in the row above. Hence, Pascal's triangle is easy to write down row by row.⁸

The entries in each row also arise in another way. The formulas

$$\begin{aligned}
 (1+x)^2 &= 1 + 2x + x^2, \\
 (1+x)^3 &= 1 + 3x + 3x^2 + x^3, \\
 (1+x)^4 &= 1 + 4x + 6x^2 + 4x^3 + x^4,
 \end{aligned}$$

are easily verified, and the coefficients on the right side in each case are the integers in rows 2, 3, and 4 of Pascal's triangle. The general result follows by induction, and will be used several times in this book.

Example 6. Prove the Binomial Theorem:

$$(1+x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + \cdots + \binom{n}{n}x^n, \quad \text{for all } n \geq 0.$$

Solution. The theorem holds if $n = 0$ because $\binom{0}{0} = 1$ and $(1+x)^0 = 1$. If it holds for some $k \geq 0$ then, using the Pascal identity, we obtain

$$\begin{aligned}
 (1+x)^{k+1} &= (1+x)(1+x)^k \\
 &= (1+x) \left[\binom{k}{0} + \binom{k}{1}x + \cdots + \binom{k}{k-1}x^{k-1} + \binom{k}{k}x^k \right] \\
 &= \binom{k}{0} + \left[\binom{k}{0} + \binom{k}{1} \right] x + \cdots + \left[\binom{k}{k-1} + \binom{k}{k} \right] x^k + \binom{k}{k}x^{k+1} \\
 &= \binom{k+1}{0} + \binom{k+1}{1}x + \cdots + \binom{k+1}{k}x^k + \binom{k+1}{k+1}x^{k+1},
 \end{aligned}$$

which completes the induction. \square

When proving inductively that statements p_m, p_{m+1}, \dots, p_k are true, the most difficult part is usually showing that $p_k \Rightarrow p_{k+1}$ for each $k \geq m$. Clearly, this task would be easier if we could assume the truth of p_m, \dots, p_{k-1} in addition to the truth of p_k when deducing p_{k+1} . This assumption leads to a useful variant of the principle of induction (in fact, it is equivalent to it).

⁸Note that this shows the binomial coefficients are all *integers*, a fact that is not clear from the definition.

Theorem 2. Principle of Strong Induction. Let m be an integer and, for each $n \geq m$, let p_n be a statement. Suppose the following conditions are satisfied.

- (1) p_m is true.
- (2) If $k \geq m$ and all of p_m, p_{m+1}, \dots, p_k are true, then p_{k+1} is also true.

Then p_n is true for every $n \geq m$.

Proof. For each $n \geq m$, let t_n be the statement that p_m, p_{m+1}, \dots, p_n are all true. Then, t_m is true by (1). If t_k is true for some $k \geq m$, then (2) implies that p_{k+1} is true, so t_{k+1} is also true. Hence, t_n is true for all $n \geq m$ by Theorem 1, so certainly p_n is true for all $n \geq m$. \blacksquare

In the next example, we use strong induction to prove an important fact about primes that would be more difficult to deduce using (ordinary) induction. Recall that a *prime number* (or *prime*) is an integer $p \geq 2$ that cannot be factored as a product of two smaller positive integers.

Example 7. Show that every integer $n \geq 2$ is a product of (one or more) primes.

Solution. This assertion is true if $n = 2$ because 2 is a prime. If $k \geq 2$, we assume inductively that $2, 3, \dots, k$ are all products of primes. To apply strong induction, we must show that $k + 1$ is a product of primes. This is clear if $k + 1$ is itself prime; otherwise, let $k + 1 = ab$, where $2 \leq a \leq k$ and $2 \leq b \leq k$. Then both a and b are products of primes by the (strong) induction hypothesis, so $k + 1 = ab$ is also a product of primes. \square

We conclude with an intuitively clear property of \mathbb{Z} that is equivalent to the principle of induction, and which is usually taken as an axiom.

Well-Ordering Principle. Every nonempty set of nonnegative integers has a smallest member.

Proof. If the principle is false, let $X \subseteq \{0, 1, 2, \dots\}$ be a nonempty set that has no smallest member. For each $n \geq 0$, let p_n be the statement " $n \notin X$." It suffices to show that p_n is true for all $n \geq 0$ —since then X is empty, contrary to our assumption. We prove this by strong induction. First, p_0 is true because if $0 \in X$, then it is the smallest member of X (because $X \subseteq \{0, 1, 2, \dots\}$). Now assume inductively that p_0, p_1, \dots, p_k are all true, so that none of $0, 1, \dots, k$ is in X . This implies that $k + 1 \notin X$ since otherwise it would be the smallest member of X . This means p_{k+1} is true, and so completes the induction. \square

The way the well-ordering principle is used can be illustrated by the following frivolous example: Suppose that we want to show that every positive integer is interesting. If this assertion were false, the set of uninteresting positive integers would be nonempty and so would contain a smallest member by the axiom. But the smallest uninteresting integer would surely be interesting—a contradiction! This technique can also be applied to *serious* situations.

For example, the well-ordering principle implies the induction principle. Indeed, let p_1, p_2, p_3, \dots be statements such that p_1 is true and $p_k \Rightarrow p_{k+1}$ for every $k \geq 1$. If $X = \{n \geq 1 \mid p_n \text{ is false}\}$, we must show that X is empty. But if not, then X has a smallest member, which leads to a contradiction. The details are in Exercise 15.

We have proved the following implications (the first is Theorem 2):

$$\text{Induction} \Rightarrow \text{Strong Induction} \Rightarrow \text{Well Ordering.}$$

Moreover, well ordering implies induction (see above), so the three principles are logically equivalent. The validity of these principles is one of the basic **Peano axioms**⁹ for the integers.

Inductive Definition

Many arguments in algebra (in fact, in mathematics generally) refer to **sequences** $a_0, a_1, a_2, a_3, \dots, a_n, \dots$ from a set A where each a_i is an element of A called the i^{th} term of the sequence. Hence $1, 2, 4, 8, 16, \dots$ are the first five terms of the sequence $a_n = 2^n$ from \mathbb{Z} . This sequence can be compactly described as follows:

$$a_0 = 1 \quad \text{and} \quad a_n = 2a_{n-1} \quad \text{for each } n \geq 1. \quad (*)$$

These conditions uniquely describe the sequence (the formula $a_n = 2^n$ for $n \geq 0$ can be proved by induction), and for this reason (*) is called an *inductive definition* of the sequence. More generally, a sequence is said to be **defined inductively** if the first term is specified and each later term is uniquely determined by the earlier terms (often by a formula). It is usually very difficult to give an explicit formula for the n^{th} term a_n in terms of the earlier terms; nevertheless, the following theorem shows that such a sequence always exists and is uniquely determined.

Theorem 3. Recursion Theorem. Given a set A and $a \in A$, there is exactly one sequence $a_0, a_1, a_2, a_3, \dots, a_n, \dots$ from A that satisfies the following requirements:

- (1) $a_0 = a$.
- (2) For each $n \geq 1$, the term a_n is uniquely determined by the preceding terms $a_0, a_1, a_2, \dots, a_{n-1}$.

Proof. The existence of such a sequence is given in Appendix D; we prove uniqueness by strong induction on $n \geq 0$. Clearly, a_0 is uniquely determined by (1). If each of $a_0, a_1, a_2, \dots, a_{n-1}$ has been uniquely specified, then a_n is uniquely determined by (2). Hence, the sequence is uniquely determined by (1) and (2). ■

Exercises 1.1

1. Prove each equation by induction on n .
 - (a) $1 + 5 + 9 + \dots + (4n - 3) = n(2n - 1)$ for all $n \geq 1$.
 - (b) $1^2 + 2^2 + \dots + n^2 = \frac{1}{6}n(n+1)(2n+1)$ for all $n \geq 1$.
 - (c) $1^3 + 2^3 + \dots + n^3 = \frac{1}{4}n^2(n+1)^2$ for all $n \geq 1$.
 - (d) $1 \cdot 2 + 2 \cdot 3 + \dots + n \cdot (n+1) = \frac{1}{3}n(n+1)(n+2)$ for all $n \geq 1$.
 - (e) $1 \cdot 2^2 + 2 \cdot 3^2 + \dots + n \cdot (n+1)^2 = \frac{1}{12}n(n+1)(n+2)(3n+5)$ for all $n \geq 1$.
 - (f) $\frac{1}{1^2} + \frac{1}{2^2} + \dots + \frac{1}{n^2} = \frac{n}{n+1}$ for all $n \geq 1$.
 - (g) $1^2 + 3^2 + \dots + (2n-1)^2 = \frac{n}{3}(4n^2 - 1)$ for all $n \geq 1$.

⁹Named after Giuseppe Peano, an Italian mathematician and logician who, in 1889, reduced the theory of the natural numbers \mathbb{N} to five simple axioms. For a discussion of this, see R.A. Beaumont and R.S. Pierce, *The Algebraic Foundations of Mathematics*, Addison-Wesley, 1963.

- (h) $1^2 - 2^2 + 3^2 - \cdots + (-1)^{n+1}n^2 = \frac{1}{2}(-1)^{n+1}n(n+1)$ for all $n \geq 1$.
 (i) $\frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \cdots + \frac{n}{(n+1)!} = 1 - \frac{1}{(n+1)!}$ for all $n \geq 1$.
2. Prove each inequality by induction on n .
- (a) $n < 2^n$ for all $n \geq 0$.
 - (b) $n^2 \leq 2^n$ for all $n \geq 4$.
 - (c) $n! \leq 2^{n^2}$ for all $n \geq 4$ (compare with Example 5).
 - (d) $\frac{1}{1^2} + \frac{1}{2^2} + \cdots + \frac{1}{n^2} \leq 2 - \frac{1}{n}$ for all $n \geq 1$.
 - (e) $\frac{1}{\sqrt{1}} + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}} \geq \sqrt{n}$ for all $n \geq 1$.
 - (f) $\frac{1}{\sqrt[3]{1}} + \frac{1}{\sqrt[3]{2}} + \cdots + \frac{1}{\sqrt[3]{n}} \leq 2\sqrt[3]{n} - 1$ for all $n \geq 1$.
3. Prove each statement by induction on n .
- (a) $n^3 + (n+1)^3 + (n+2)^3$ is a multiple of 9 for all $n \geq 1$.
 - (b) $n^3 - n$ is a multiple of 3 for all $n \geq 1$.
 - (c) $3^{2n+1} + 2^{n+2}$ is a multiple of 7 for all $n \geq 0$.
4. Show that $(1 - \frac{1}{2^2})(1 - \frac{1}{3^2}) \cdots (1 - \frac{1}{n^2}) = \frac{n+1}{2n}$ for all $n > 2$.
5. Show that $3^{3n} + 1$ is a multiple of 7 for all odd $n \geq 1$.
6. Suppose that n straight lines in the plane are positioned so that no two are parallel and no three pass through the same point. Show that they divide the plane into $\frac{1}{2}(n^2 + n + 2)$ distinct regions.
7. Show that there are 3^n positive integers with n digits, where each digit must be 4, 5, or 6.
8. A polygon in the plane is called *convex* if every line joining two vertices is either an edge or lies entirely within the polygon. If $n \geq 3$, show that the sum of the interior angles of an n -sided convex polygon equals $(n - 2) \cdot 180^\circ$.
9. A straight line segment joining two distinct points on a circle is called a *secant*. For $n \geq 1$, draw n secants with no two identical. Show that the resulting regions can be unambiguously colored black and white (where *unambiguously* means that no two regions sharing a straight line boundary are of the same color).
10. (a) Show that any postage of $n \geq 2$ cents can be made of 2 and 3 cent stamps.
 (b) Show that any postage of $n \geq 12$ cents can be made of 3 and 7 cent stamps.
 (c) Show that any postage of $n \geq 18$ cents can be made of 4 and 7 cent stamps.
 (d) Can you generalize from the results in (a)–(c)?
11. Let $a_n = 2^{3n} - 1$ for $n \geq 0$. Guess a common divisor of each a_n and prove your assertion.
12. (a) Try to prove the statement “ $1^3 + 2^3 + \cdots + n^3$ is a perfect square” by induction. Now look at Exercise 1(c).
 (b) Try to prove that $1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n} < 2$ by induction. Now formulate a stronger equality for the sum on the left, prove it by induction, and use it to deduce the inequality.
13. Prove the **Pascal identity**: $\binom{n}{r-1} + \binom{n}{r} = \binom{n+1}{r}$ for $1 \leq r \leq n$.
14. (a) Show that $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$ for all $n \geq 0$.
 (b) Show that $\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \cdots \pm \binom{n}{n} = 0$ if $n > 0$.
15. Use the well-ordering principle to prove the principle of induction. [Hint: See the discussion following the well-ordering principle.]
16. Let X be a nonempty set of integers. Then X is said to be *bounded below* (*bounded above*) if an integer m exists such that $m \leq x$ for all $x \in X$ (respectively $m \geq x$ for all $x \in X$).

- (a) If X is bounded below, show that it has a smallest member.
 (b) If X is bounded above, show that it has a largest member.
17. Use strong induction to prove that every integer $n \geq 2$ has a prime factor.
18. In each case, conjecture a formula for a_n and prove it by induction.
- $a_0 = 2, a_{n+1} = -a_n, n \geq 0.$
 - $a_0 = 1, a_1 = -2, a_{n+2} = 2a_n - a_{n+1}, n \geq 0.$
 - $a_0 = 1, a_{n+1} = 1 - a_n, n \geq 0.$
 - $a_0 = 3, a_{n+1} = (a_n)^2, n \geq 0.$
19. Let n lines in the plane be such that no two are parallel and no three are concurrent. Find the number a_n of regions into which the plane is divided by first showing that $a_{n+1} = a_n + (n + 1)$.
20. Prove the following induction principle.
 Let m be an integer and let p_n be a statement for all $n \geq m$. Assume that
 (1) p_m and p_{m+1} are true.
 (2) If $k \geq m$ and both p_k and p_{k+1} are true, then p_{k+2} is true.
 Then p_n is true for all $n \geq m$.
21. Let a_n denote a number for each integer $n \geq 0$ and assume that $a_{n+2} = a_{n+1} + 2a_n$ holds for every $n \geq 0$. Use the principle in Exercise 20 to prove each assertion.
- If $a_0 = 1$ and $a_1 = -1$, then $a_n = (-1)^n$ for each $n \geq 0$.
 - If $a_0 = 1$ and $a_1 = 2$, then $a_n = 2^n$ for each $n \geq 0$.
 - If $a_0 = p$ and $a_1 = q$, then $a_n = \frac{1}{3}[(p+q)2^n + (2p-q)(-1)^n]$ for each $n \geq 0$.
22. Let p_n denote the statement: “ $3n + 2$ is a multiple of 3.” Show that $p_k \Rightarrow p_{k+1}$ for all $k \geq 1$. What does this say about Theorem 1?
23. Let p_n denote the statement: “In any class of n algebra students, every student obtains the same grade.” Then p_1 is clearly true. If p_n is satisfied for $n > 1$, suppose that x_1, x_2, \dots, x_{n+1} denotes a class of $n + 1$ students. Then x_1, x_2, \dots, x_n all have the same grade (by induction) as do x_2, x_3, \dots, x_{n+1} . Thus x_1, x_2, \dots, x_{n+1} all have the same grade (the same as x_n), so p_{n+1} is true. Hence, p_n is true for all n . What is wrong with this argument?
24. Suppose that p_n is a statement about n for each $n \geq 1$. In each case what must be done to prove that p_n is true for all $n \geq 1$?
- $p_n \Rightarrow p_{n+2}$ for each $n \geq 1$.
 - $p_n \Rightarrow p_{n+8}$ for each $n \geq 1$.
 - $p_n \Rightarrow p_{n+1}$ for each $n \geq 10$.
25. If p_n is a statement about n for each $n \geq 1$, argue that p_n is true for all $n \geq 1$ if $p_n \Rightarrow p_{n-1}$ for each $n \geq 2$ and p_n is true for infinitely many values of n .
26. For a sequence a_1, a_2, \dots , suppose that $a_1 + a_2 + \dots + a_n$ is to be evaluated.
- If a sequence b_1, b_2, \dots can be found such that $a_n = b_{n+1} - b_n$ for all $n > 1$, prove by induction that $a_1 + a_2 + \dots + a_n = b_{n+1} - b_1$.
 - Use the technique in (a) to evaluate $1 \cdot 2 \cdot 3 + 2 \cdot 3 \cdot 4 + \dots + n(n+1)(n+2)$.
 [Hint: Try $b_n = (n-1)n(n+1)(n+2)$.]
27. Suppose that a sequence a_0, a_1, \dots is given.
- Show that the sequence s_0, s_1, \dots exists where $s_0 = a_0$ and s_n is the sum of the first $n + 1$ of a_i .
 - Show that the sequence p_0, p_1, \dots exists where $p_0 = a_0$ and p_n is the product of the first $n + 1$ of the a_i .

1.2 DIVISORS AND PRIME FACTORIZATION

Mathematics is the queen of the sciences and number theory is the queen of mathematics.

—Carl Friedrich Gauss

The set \mathbb{Z} of integers will be used in several ways throughout this book: as a major source of examples of algebraic systems; to state definitions and prove theorems (often by induction); and as a prototype for results about more general systems. For the most part, the properties of \mathbb{Z} that we need are familiar facts about addition, multiplication, and ordering of the integers, although we present a more detailed look at these properties in Section 3.2. However, we also utilize several less familiar properties of divisibility and primes in \mathbb{Z} and so devote this section to them.

The Greatest Common Divisor

When we write $22/7$ in the form $3\frac{1}{7}$ we are using the fact that $22 = 3 \cdot 7 + 1$; that is, 22 leaves a remainder of 1 when divided by 7. The general result is a consequence of the well-ordering axiom.

Theorem 1. Division Algorithm. Let n and $d \geq 1$ be integers. There exist uniquely determined integers q and r such that

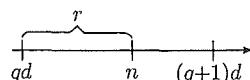
$$n = qd + r \quad \text{and} \quad 0 \leq r < d.$$

Proof. Let $X = \{n - td \mid t \in \mathbb{Z}, n - td \geq 0\}$. Then X is nonempty. In fact, if $n \geq 0$, then $n = n - 0d$ is in X ; if $n < 0$, then $n - nd = n(1 - d)$ is in X . Hence, by the well-ordering principle, let r be the smallest member of X . Then $r = n - qd$ for some q and $r \geq 0$, so it remains to show that $r < d$. But if $r \geq d$, then $0 \leq r - d = n - (q+1)d$. This means that $r - d$ is in X , contradicting the minimality of r . This result proves the existence of q and r .

To prove uniqueness, suppose also that $n = q'd + r'$ with $0 \leq r' < d$. Assume $r \leq r'$ (the case $r' \leq r$ is similar). Then $(q - q')d = r' - r$ is a nonnegative, integral multiple of d that is less than d (because $r' - r \leq r' < d$). This can occur only if $r = r'$, which implies that $q = q'$ and so proves uniqueness. ■

For n and $d \geq 1$, the integers q and r in Theorem 1 are called the **quotient** and **remainder**, respectively. Thus, for example, if we divide $n = -17$ by $d = 5$, the result is $-17 = (-4) \cdot 5 + 3$, so the quotient is -4 and the remainder is 3.

The division algorithm can also be seen geometrically. If the real line is marked off in multiples of d , n clearly falls either on a multiple qd of d or between qd and $(q+1)d$ (see the diagram). Hence, $qd \leq n < (q+1)d$, so $0 \leq n - qd < d$, and we take $r = n - qd$.



If both n and d are positive and a calculator is available, the quotient q and the remainder r can be easily found as follows: Calculate $\frac{n}{d}$ and let q denote the largest integer that is less than or equal to $\frac{n}{d}$. Hence,

$$0 \leq \frac{n}{d} - q < 1.$$

If we multiply through by d , we get $0 \leq n - qd < d$, so take $r = n - qd$.

Example 1. Find the quotient and remainder if $n = 4187$ and $d = 129$.

Solution. We have $\frac{n}{d} = 32.457$ approximately, so $q = 32$. Then $r = n - dq = 59$, and so $4187 = 32 \cdot 129 + 59$, as desired. \square

If n and d are integers, d is called a **divisor** of n if $n = qd$ for some integer q . When this is the case, we write $d|n$. If $d|n$ is not true, we write $d \nmid n$. Thus, $7|84$ but $7 \nmid 85$. Note that $1|n$ and $n|0$ for all integers n . The following properties of divisors will be used frequently.

Theorem 2. Let m, n and d denote integers.

- (1) $n|n$ for all n .
- (2) If $d|m$ and $m|n$, then $d|n$.
- (3) If $d|n$ and $n|d$, then $d = \pm n$.
- (4) If $d|n$ and $d|m$, then $d|(xn + ym)$ for all integers x and y .

Proof. The proofs of (1) and (2) are left to the reader. In (3), let $n = qd$ and $d = pn$ for integers p and q . If $d = 0$, then $n = qd = 0 = d$. If $d \neq 0$, then $d = pn = pqd$, which implies that $1 = pq$. As p and q are integers, this means that $p = q = 1$ or $p = q = -1$, and so $d = n$ or $d = -n$, which proves (3). As to (4), if $n = ad$ and $m = bd$ in (4), then $xn + ym = (xa + yb)d$, so $d|(xn + ym)$, as required. \blacksquare

Expressions of the form $xn + ym$, where x and y are integers, are called **linear combinations** of n and m .

Example 2. If $d \geq 1$ is such that $d|(3k + 5)$ and $d | (7k + 2)$ for some k , show that $d = 1$ or $d = 29$.

Solution. The hypotheses and (4) of Theorem 2 imply that d divides the linear combination $7(3k + 5) - 3(7k + 2) = 35 - 6 = 29$. Hence, d is a positive divisor of 29, so $d = 1$ or $d = 29$. \square

An integer d is called a **common divisor** of two integers m and n if $d|m$ and $d|n$. To motivate the next theorem, consider the positive divisors of 36 and 84:

- Positive divisors of 36: 1, 2, 3, 4, 6, 9, 12, 18, 36
- Positive divisors of 84: 1, 2, 3, 4, 6, 7, 12, 14, 21, 28, 42, 84
- Common divisors: 1, 2, 3, 4, 6, 12

We wish to focus attention on the fact that the largest common divisor 12 is actually a *multiple* of all the other positive common divisors. This idea is built into the following definition. Let m and n be integers.

An integer d is called a **greatest common divisor** of m and n if:

- (1) $d \geq 1$
- (2) $d|m$ and $d|n$
- (3) If $k|m$ and $k|n$, then $k|d$.

When it exists we write $d = \gcd(m, n)$.

For example, $\gcd(18, 30) = 6$, $\gcd(6, 7) = 1$, and $\gcd(-9, 15) = 3$.

Conditions (2) and (3) can be stated as follows: $\gcd(m, n)$ is a common divisor of m and n by (2), which is a multiple of every common divisor by (3). If it exists,

$d = \gcd(m, n)$ is unique. In fact, if d' is another integer satisfying (1), (2), and (3), then $d'|d$ by (3). Similarly, $d|d'$ so $d = \pm d'$ by Theorem 2. But then $d' = d$ because we insist that greatest common divisors are positive.

The following fundamental theorem shows that, if m and n are not both zero, then $d = \gcd(m, n)$ does indeed exist and, surprisingly, that d is actually a linear combination of m and n .

Theorem 3. *Let m and n be integers, not both zero. Then $d = \gcd(m, n)$ exists and $d = xm + yn$ for some integers x and y .*

Proof. Let $X = \{xm + yn \mid x, y \in \mathbb{Z}, xm + yn \geq 1\}$. Then X is not empty because $m^2 + n^2 \in X$, so let d be the smallest member of X (by the well-ordering principle). Since $d \in X$, we have $d \geq 1$ and $d = xm + yn$ for integers x and y . Also, if $k|m$ and $k|n$, then $k|(xm + yn) = d$ by Theorem 2. So it remains to show that $d|m$ and $d|n$.

To show that $d|m$, write $m = qd + r$ where $0 \leq r \leq d - 1$. Then,

$$r = m - qd = m - q(xm + yn) = (1 - qx)m + (-qy)n.$$

Hence, if $r \geq 1$, then $r \in X$ and $r < d$, contradicting the choice of d . So $r = 0$, that is, $m = qd$. Thus, $d|m$, and $d|n$ is proved similarly. \blacksquare

Note that $\gcd(m, n)$ does *not* exist if $m = 0 = n$ (verify), which explains the requirement in Theorem 3 that m and n are not both zero. Also, the greatest common divisor of m and n can be a linear combination of m and n in more than one way. For example, $\gcd(2, 3) = 1$ and we have $1 = 2 \cdot 1 - 3$ and $1 = 3 - 2$.

Example 3. If p and q are distinct primes, show that $\gcd(p, q) = 1$.

Solution. Write $d = \gcd(m, n)$. Then $d|p$, so $d = 1$ or p . Similarly, $d = 1$ or q , so $d = 1$ because, otherwise, $p = d = q$ is contrary to the assumption that $p \neq q$. \square

The next example (which is needed later) illustrates how the definition of the greatest common divisor is used.

Example 4. If $m = qn + r$, show that $\gcd(m, n) = \gcd(n, r)$.

Solution. Write $d = \gcd(m, n)$ and $k = \gcd(n, r)$. Then k divides both n and r and so divides $m = qn + r$. Thus, k is a common divisor of m and n , so $k|d$ because $d = \gcd(m, n)$. A similar argument (using $r = -qn + m$) shows that $d|k$, so $d = \pm k$ by (3) of Theorem 2. Hence, $d = k$, because both d and k are positive. \square

How do we compute $d = \gcd(m, n)$ in general? There is an efficient procedure for doing so, which also shows how to express d as a linear combination of m and n . To illustrate how it works, consider the numbers 78 and 30. The idea is to use the division algorithm repeatedly. First divide 78 by 30:

$$\begin{aligned} 78 &= 2 \cdot 30 + 18 \\ 30 &= 1 \cdot 18 + 12 \\ 18 &= 1 \cdot 12 + 6 \\ 12 &= 2 \cdot 6 + 0 \end{aligned}$$

At each stage (after the first) we divide the divisor at the previous stage by the remainder at that stage. The last nonzero remainder is 6, and this equals $\gcd(78, 30)$.

This is no coincidence as we shall see. To express 6 as a linear combination of 78 and 30, eliminate the remainders from the second last lineup:

$$\begin{aligned} 6 &= 18 - 1 \cdot 12 \\ &= 18 - (30 - 1 \cdot 18) \\ &= 2 \cdot 18 - 30 \\ &= 2(78 - 2 \cdot 30) - 30 \\ &= 2 \cdot 78 - 5 \cdot 30 \end{aligned}$$

This procedure is called the **euclidean algorithm**, and it works in general. For positive integers m and n , not both zero, we use the division algorithm repeatedly:

$$\begin{array}{ll} m = q_1 n + r_1 & r_1 < n \\ n = q_2 r_1 + r_2 & r_2 < r_1 \\ r_1 = q_3 r_2 + r_3 & r_3 < r_2 \\ \vdots & \vdots \end{array}$$

At each stage we divide the divisor at the previous stage by the remainder, so the remainders form a decreasing sequence of nonnegative integers:

$$n > r_1 > r_2 > r_3 > \cdots \geq 0.$$

Clearly, we must encounter a remainder of 0 (in at most n steps). If r_t denotes the last nonzero remainder, the last two equations are

$$r_{t-2} = q_t r_{t-1} + r_t \quad \text{and} \quad r_{t-1} = q_{t+1} r_t + 0.$$

Now, repeated application of the result in Example 4 gives

$$\gcd(m, n) = \gcd(n, r_1) = \gcd(r_1, r_2) = \cdots = \gcd(r_{t-1}, r_t) = r_t.$$

Hence, $\gcd(m, n)$ really is the last nonzero remainder.

Example 5. Find $\gcd(41, 12)$ and express it as a linear combination of 41 and 12.

Solution. The algorithm is not needed to find $\gcd(41, 12)$. In fact, 1 and 41 are the only positive divisors of 41, so $\gcd(41, 12) = 1$ because 41 does not divide 12. However, guessing a linear combination $1 = x \cdot 41 + y \cdot 12$ is not easy. The euclidean algorithm gives

$$\begin{aligned} 41 &= 3 \cdot 12 + 5 \\ 12 &= 2 \cdot 5 + 2 \\ 5 &= 2 \cdot 2 + 1 \\ 2 &= 2 \cdot 1 + 0 \end{aligned}$$

Hence, $\gcd(41, 12) = 1$ as expected. Elimination of remainders gives

$$\begin{aligned} 1 &= 5 - 2 \cdot 2 \\ &= 5 - 2(12 - 2 \cdot 5) \\ &= 5 \cdot 5 - 2 \cdot 12 \\ &= 5(41 - 3 \cdot 12) - 2 \cdot 12 \\ &= 5 \cdot 41 - 17 \cdot 12 \end{aligned}$$

which is the required linear combination. □

The following definition will be used frequently throughout this book.

Two integers m and n are called **relatively prime** if $\gcd(m, n) = 1$.

For example, 2 and 3 are relatively prime, as are 20 and 9. Note that 1 is relatively prime to every integer n . The condition in Theorem 4 is useful.

Theorem 4. *Let m and n be integers, not both zero. Then m and n are relatively prime if and only if $1 = xm + yn$ for some integers x and y .*

Proof. If $\gcd(m, n) = 1$, then $1 = xm + yn$ by Theorem 3. Conversely, if $1 = xm + yn$, then any common divisor of m and n must divide 1. In particular, $\gcd(m, n) = 1$. ■

For example, any two consecutive integers k and $k + 1$ are relatively prime because $(k + 1) - k = 1$. Similarly, $5(6k + 5) - 6(5k + 4) = 1$ shows that 6 $k + 5$ and 5 $k + 4$ are relatively prime for any integer k .

Corollary. *If $d = \gcd(m, n)$, $m, n \in \mathbb{Z}$, then $\frac{m}{d}$ and $\frac{n}{d}$ are relatively prime.*

Proof. If $d = xm + yn$, $x, y \in \mathbb{Z}$, dividing by d gives $1 = x\frac{m}{d} + y\frac{n}{d}$. □

The following theorem contains two very useful properties of relatively prime integers, and will be referred to several times below.

Theorem 5. *Let m and n be relatively prime integers.*

- (1) *If $m|k$ and $n|k$ for some integer k , then $mn|k$.*
- (2) *If $m|kn$ for some integer k , then $m|k$.*

Proof. We first prove (1). By Theorem 4, let $1 = xm + yn$, where x and y are integers. If $k = qm$ and $k = pn$ where p and q are integers, then

$$k = 1 \cdot k = xmk + ynk = xm(pn) + yn(qm) = (xp + yq)m.$$

Hence, $mn|k$, proving (1). As to (2), let $nk = qm$ where q is an integer. Then,

$$k = 1 \cdot k = xmk + ynk = xmk + y(qm) = (xk + yq)m.$$

This shows that $m|k$, and so proves (2). ■

Prime Factorization

Clearly, every integer $n \geq 2$ has at least two positive divisors: 1 and n . The integers for which these are the *only* positive divisors are important. An integer p is called a **prime** if it satisfies the following conditions:

- (1) $p \geq 2$.
- (2) *If $d|p$ and $d > 0$, then either $d = 1$ or $d = p$.*

Thus, the first few primes are 2, 3, 5, 7, 11, 13, We know (Example 7 §1.1) that every integer greater than 1 is a product of primes; the reason for not regarding 1 as a prime is to ensure that this factorization is unique (see Theorem 7).

If the product of two integers is even, one of these integers must be even (because the product of two odd integers is odd). We can rephrase this statement as follows: If $2|m n$, where m and n are integers, then $2|m$ or $2|n$. This statement holds for any prime in place of 2.

Theorem 6. Euclid's Lemma. Let p denote a prime.

- (1) If $p|mn$ where m and n are integers, then $p|m$ or $p|n$.
- (2) If $p|m_1m_2\cdots m_r$ where each m_i is an integer, then $p|m_i$ for some i .

Proof. (1) Write $d = \gcd(m, p)$. Then $d|p$, so $d = 1$ or $d = p$ because p is a prime. If $d = p$, then $p|m$ because $d|m$; if $d = 1$, then $p|n$ by (2) of Theorem 5.

(2) This assertion follows by induction on r . If $r = 1$, it is obvious. If (2) holds for some $r \geq 1$, let $p|m_1m_2\cdots m_r m_{r+1}$. Then (1) shows that either $p|m_1\cdots m_r$ or $p|m_{r+1}$. In the first case, $p|m_i$ for some $i = 1, 2, \dots, r$ by the induction hypothesis. Hence, in any case, $p|m_i$ for some $i = 1, 2, \dots, r + 1$, completing the induction. ■

Note that Euclid's lemma fails for nonprimes. For example, 6 is a divisor of $3 \cdot 4$, but 6 does not divide 3 or 4.

It is not too difficult to convince yourself that every integer $n \geq 2$ is either a prime itself or can be factored as a product of primes—just keep factoring as long as possible. For example, $12 = 2^2 \cdot 3$, $25 = 5^2$, and $360 = 2^3 \cdot 3^2 \cdot 5$. In fact, *every* integer greater than 1 is a product of primes, and this factorization is unique up to the order of the factors.

Theorem 7. Prime Factorization Theorem.

- (1) Every integer $n \geq 2$ is a product of (one or more) primes.
- (2) This factorization is unique up to the order of the factors. That is, if

$$n = p_1p_2\cdots p_r \quad \text{and} \quad n = q_1q_2\cdots q_s,$$

where p_i and q_j are primes, then $r = s$ and q_j can be relabeled so that $p_i = q_i$ for all $i = 1, 2, \dots, r$.

Proof. We proved (1) in Example 7 §1.1. If (2) fails, let (by the well-ordering principle) $m \geq 2$ be the smallest integer with two distinct factorizations into primes:

$$m = p_1p_2\cdots p_r = q_1q_2\cdots q_s.$$

Then m is not a prime (verify), so $r \geq 2$ and $s \geq 2$. We have $p_1|q_1q_2\cdots q_s$, so $p_1|q_j$ for some j by Euclid's lemma. By relabeling q_j , we may assume that $p_1|q_1$. Then $p_1 = q_1$ because both are primes, so

$$\frac{m}{p_1} = p_2\cdots p_r = q_2\cdots q_s$$

is an integer—smaller than m —that admits two distinct factorizations into primes. This result contradicts the choice of m , and so proves (2). ■

Corollary. Two integers $m \geq 2$ and $n \geq 2$ are relatively prime if and only if no prime divides both m and n .

Proof. Write $d = \gcd(m, n)$. If $d = 1$, then any common prime divisor would have to divide 1, so no such common divisor exists. Conversely, suppose no prime divides both m and n . If $d > 1$ and $p|d$ where p is a prime, then $p|m$ and $p|n$, contrary to our assumption. So $d = 1$, that is m and n are relatively prime. □

If $n \geq 2$ is an integer and p_1, p_2, \dots, p_r are the distinct prime divisors of n , the prime factorization theorem asserts that n can be written uniquely in the form

$$n = p_1^{n_1}p_2^{n_2}\cdots p_r^{n_r},$$

where $n_i \geq 1$ for each i . This means that the primes p_i and the integers n_i are uniquely determined by n . For example, $60 = 2^2 \cdot 3 \cdot 5$ and $882 = 2 \cdot 3^2 \cdot 7^2$.

If n has only one prime divisor, we call it a **prime power**, examples being $7 = 7^1$, $9 = 3^2$, and $32 = 2^5$. At the other extreme, we say that n is **square free** if all the exponents $n_i = 1$. Hence, any prime is square free as are $6 = 2 \cdot 3$ and $70 = 2 \cdot 5 \cdot 7$.

If n is not prime, it must have a prime divisor $p \leq \sqrt{n}$ (it cannot have two prime divisors greater than \sqrt{n}). So to test whether n is prime, it suffices to verify that it has no prime divisor $p \leq \sqrt{n}$ (which is impractical if n is very large).

Example 6. Factor 1591 into primes.

Solution. We start dividing 1591 by the successive primes, 2, 3, 5, 7, Since $\sqrt{1591} < 40$ (because $40^2 = 1600$), we need go only as high as 37; in fact, the first prime that divides 1591 is 37. As $1591 = 37 \cdot 43$ and 43 is a prime, we have the required prime factorization. \square

Obviously, the method in Example 6 requires that we have a list of the primes. Although large tables of primes are available, the method clearly fails for very large numbers. Finding the prime factorization of large integers is very difficult. Even so, on December 15, 2005 it was announced that $2^{30,402,457} - 1$ is a prime with 9,152,052 digits, the largest prime known to that date. Such a result requires a very large amount of computer time.¹⁰

The prime factorization theorem gives a systematic way of listing all the positive divisors of an integer n when the prime factorization of n is known. For example, if $n = 12 = 2^3 \cdot 3$, these divisors are 1, 2, 3, 4, 6, and 12, and they can be written as

$$\begin{aligned} 1 &= 2^0 3^0 & 2 &= 2^1 3^0 & 4 &= 2^2 3^0 \\ 3 &= 2^0 3^1 & 6 &= 2^1 3^1 & 12 &= 2^2 3^1 \end{aligned}$$

Thus, they can all be expressed as $2^r 3^s$, where $0 \leq r \leq 2$ and $0 \leq s \leq 1$ (where $p^0 = 1$ for any prime p). The general situation is as follows:

Theorem 8. Let n be an integer with prime factorization

$$n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r},$$

where p_i are distinct primes and $n_i \geq 1$ for each i . Then the positive divisors of n are precisely the integers d of the form:

$$d = p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r},$$

where $0 \leq d_i \leq n_i$ holds for each i .

Proof. The prime divisors of d are contained in $\{p_1, \dots, p_r\}$ by Euclid's lemma, and d cannot contain a higher power of p_i than $p_i^{n_i}$ by Theorem 7. \blacksquare

In much the same way, the prime factorization theorem provides a simple way to compute the greatest common divisor of any finite set of positive integers (rather

¹⁰On the other hand, in 2002, Manindra Agrawal and two undergraduate students (Neeraj Kayal and Nitin Saxena) gave a simple algorithm that can decide whether a given integer n is prime or not. Moreover, the time taken is approximately a polynomial function of n . This is an important breakthrough in computer science.

than just two). It also provides the “dual” notion, the least common multiple. The definitions are as follows. Let n_1, n_2, \dots, n_r be positive integers.

- (1) **The greatest common divisor** $\gcd(n_1, n_2, \dots, n_r)$ of these integers is the positive common divisor that is a multiple of every common divisor.
- (2) **The least common multiple** $\text{lcm}(n_1, n_2, \dots, n_r)$ of these integers is the positive common multiple that is a divisor of every common multiple.

Thus, $\gcd(4, 6, 10) = 2$ and $\text{lcm}(4, 6, 10) = 60$ by inspection. Theorem 9 below shows that the gcd and lcm always exist. They are uniquely determined in the same way as the gcd of two integers (see the discussion preceding Theorem 3). The next example illustrates a systematic method for finding the gcd and lcm.

Example 7. Find $d = \gcd(12, 20, 18)$ and $m = \text{lcm}(12, 20, 18)$.

Solution. We might find $d = 2$ by experiment, but $m = 180$ is not clear. A systematic method involves writing the prime factorizations as follows:

$$12 = 2^2 \cdot 3^1 \cdot 5^0$$

$$20 = 2^2 \cdot 3^0 \cdot 5^1$$

$$18 = 2^1 \cdot 3^2 \cdot 5^0$$

We have $d = 2^a \cdot 3^b \cdot 5^c$ for some a, b , and c by Theorem 8. We have $a \leq 1$ because $d|18$, and $b = c = 0$ because $d|20$ and $d|12$. Thus, $d = 2$ is the largest possibility. Similarly, write the prime factorization of m as $m = 2^p \cdot 3^q \cdot 5^r \cdot k$, where $k \geq 1$ is the factor involving primes (if any) other than 2, 3, or 5. Then $p \geq 2$ because $12|m$ (or because $20|m$), $q \geq 2$ because $18|m$, and $r \geq 1$ because $20|m$. The smallest possibility is thus $m = 2^2 \cdot 3^2 \cdot 5^1 = 180$. \square

In Example 7, the power of 2 in $d = \gcd(12, 20, 18)$ is the *smallest* of the powers of 2 occurring in 12, 20, and 18; the same is true for the powers of 3 and 5 in d . Similarly, the power of 2 in $m = \text{lcm}(12, 20, 18)$ is the *largest* of the powers of 2 in 12, 20, and 18, with similar statements for the primes 3 and 5. This method works in general. For finitely many integers a, b, c, \dots , let

$$\max(a, b, c, \dots) \quad \text{and} \quad \min(a, b, c, \dots)$$

denote the largest and the smallest of these integers, respectively. For example, we have $\max(3, 1, -5, 3) = 3$ and $\min(1, 0, 5) = 0$.

Using Theorem 8, the solution to Example 7 extends to a proof of Theorem 9.

Theorem 9. Let $\{a, b, c, \dots\}$ be a finite set of positive integers, and write

$$a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$$

$$b = p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r}$$

$$c = p_1^{c_1} p_2^{c_2} \cdots p_r^{c_r}$$

$$\vdots$$

where p_i are primes dividing at least one of a, b, c, \dots , and where an exponent is zero if the prime in question does not occur in that number. Then,

$$\gcd(a, b, c, \dots) = p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r},$$

$$\text{lcm}(a, b, c, \dots) = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r},$$

where $k_i = \min(a_i, b_i, c_i, \dots)$ and $m_i = \max(a_i, b_i, c_i, \dots)$ for each i .

Example 8. Find $\gcd(63, 60, 105)$ and $\text{lcm}(63, 60, 105)$.

Solution. The prime factorizations are

$$63 = 2^0 3^2 5^0 7^1, \quad 60 = 2^2 3^1 5^1 7^0, \quad \text{and} \quad 105 = 2^0 3^1 5^1 7^1.$$

Hence, $\gcd(63, 60, 105) = 2^0 3^1 5^0 7^0 = 3$ and $\text{lcm}(63, 60, 105) = 2^2 3^2 5^1 7^1 = 1260$. \square

Of course we can use Theorem 9 to find $\text{lcm}(a, b)$ and $\gcd(a, b)$ for two integers a and b . However, the Euclidean algorithm is also available to compute $\gcd(a, b)$, so the next result is useful for finding $\text{lcm}(a, b)$.

Corollary. If a and b are positive integers, then $\text{lcm}(a, b) \cdot \gcd(a, b) = ab$.

Proof. The assertion follows from Theorem 9 and the fact that, for integers m and n , $\max(m, n) + \min(m, n) = m + n$. \blacksquare

Note that $\text{lcm}(a, b, c) \cdot \gcd(a, b, c) \neq abc$ can occur (consider Example 8).

We conclude with one last application of the prime factorization theorem.

Theorem 10. Euclid's Theorem. There are infinitely many primes.

Proof. Suppose, on the contrary, that there are only n primes, denoted p_1, p_2, \dots, p_n . Then consider the integer $m = 1 + p_1 p_2 \cdots p_n$. Since $m \geq 2$, some prime divides m by Theorem 7. But if $p_i | m$, then p_i divides $m - p_1 p_2 \cdots p_m = 1$, a contradiction. Hence the assumption that there are only finitely many primes is untenable. \blacksquare

Euclid's theorem certainly implies that there are infinitely many odd primes, that is, primes of the form $2k + 1$, $k = 0, 1, \dots$, and a natural question is whether there are infinitely many primes of the form $mk + n$ for any positive integers m and n . This clearly cannot happen unless m and n are relatively prime. However, in this case it is valid, a result first proved by P.G.L. Dirichlet. One instance of Dirichlet's theorem is treated in Exercise 39.

However, there are many unanswered questions about primes, among them the celebrated **Goldbach conjecture**, which asserts that every even integer greater than 2 is the sum of two primes. The conjecture dates from 1742 and originated in some correspondence between C. Goldbach and L. Euler. It is not known whether this assertion is true; the question appears to be extremely difficult to answer. The best result known is that every sufficiently large even number is the sum of a prime and a number that is the product of at most two primes.

Exercises 1.2

1. In each case find the quotient and remainder when n is divided by d .
 - (a) $n = 391$, $d = 17$
 - (b) $n = 401$, $d = 19$
 - (c) $n = -116$, $d = 13$
 - (d) $n = -162$, $d = 17$
2. In each case write $r = n - qd$, as in Example 1.
 - (a) $n = 51837$, $d = 386$
 - (b) $n = 39214$, $d = 871$
3. If n and $d \neq 0$ are integers, show that integers q and r exist such that $n = qd + r$ and $0 \leq r < |d|$.
4. Show that the negative divisors of an integer n are just the negatives of the positive divisors.

5. If m and n are odd integers, show that $m^2 - n^2$ is divisible by 8.
6. Given three consecutive integers, show that one must be a multiple of 3.
7. (a) If $d > 0$, $d|(11k + 4)$, and $d|(10k + 3)$ for some integer k , show that $d = 1$ or $d = 7$.
 (b) If $d > 0$, $d|(35k + 26)$, and $d|(7k + 3)$ for some integer k , show that $d = 1$ or $d = 11$.
8. Explain why $\gcd(0, 0)$ does not exist. If $n > 0$, what is $\gcd(0, n)$?
9. In each case, compute $\gcd(m, n)$ and express it as a linear combination of m and n .

(a) $m = 72, n = 42$	(b) $m = 41, n = 25$
(c) $m = 327, n = 54$	(d) $m = 198, n = 241$
(e) $m = 377, n = 29$	(f) $m = 527, n = 31$
(g) $m = 72, n = -175$	(h) $m = -231, n = 150$
10. If $m \geq 1$, show that $m|n$ if and only if $\gcd(m, n) = m$.
11. Let $d = \gcd(m, n)$. If $k|d$, $k \geq 1$, show that $\gcd(\frac{m}{k}, \frac{n}{k}) = \frac{d}{k}$.
12. If m and n are relatively prime and $k|m$, show that k and n are relatively prime.
13. Is $n^2 + n + 11$ prime for all $n \geq 1$? Support your answer.
14. Show that $\gcd(m + n, m) = \gcd(m, n)$.
15. If $m|m_1$ and $n|n_1$, show that $\gcd(m, n)|\gcd(m_1, n_1)$.
16. If $n|k(n + 1)$, show that $n|k$.
17. If $\gcd(m, n) = 1$ and $\gcd(k, n) = 1$, show that $\gcd(mk, n) = 1$.
18. If $\gcd(m, n) = 1$, let $d = \gcd(m + n, m - n)$. Show that $d = 1$ or $d = 2$.
19. Show that $\gcd(km, kn) = k \gcd(m, n)$ if $k \geq 1$.
20. Show that m and n are relatively prime if and only if no prime divides both.
21. Suppose that $p \geq 2$ is an integer with the following property: If m and n are integers and $p|mn$, either $p|m$ or $p|n$. Show that p must be a prime.
22. If d_1, \dots, d_r are all divisors of n and if $\gcd(d_i, d_j) = 1$ whenever $i \neq j$, show that $d_1 d_2 \cdots d_r$ divides n .
23. If $d = \gcd(a, n)$, must $\frac{a}{d}$ and n be relatively prime? Prove or disprove.
24. Show that any two consecutive odd integers are relatively prime.
25. Show that 3, 5, and 7 is the only *prime triple* (that is, three consecutive odd integers, each of which is prime). It is not known if there are infinitely many *prime pairs*.
26. Let p be a prime. If n is any integer, show that either $p|n$ or $\gcd(p, n) = 1$.
27. If $\gcd(m, p) = 1$ and p is a prime, show that $\gcd(m, p^k) = 1$ for all $k \geq 1$.
28. Show that none of $n! + 2, n! + 3, \dots, n! + n$ are primes for any $n \geq 2$. Hence, show that there are arbitrarily long gaps in the primes.
29. Let $ab = a_1 b_1$, where a, b, a_1 , and b_1 are positive integers. If $\gcd(a, b_1) = 1$ and $\gcd(a_1, b) = 1$, show that $a = a_1$ and $b = b_1$.
30. Find the prime factorizations of the following integers:

(a) 27783	(b) 1331	(c) 2431
(d) 18900	(e) 241	(f) 1457
31. Find the gcd and the lcm of the following pairs of numbers:

(a) 735, 110	(b) 101, 113	(c) 139, 278	(d) 221, 187
--------------	--------------	--------------	--------------
32. If $d = \gcd(a, b)$ and $m = ab/d$, show that $m = \text{lcm}(a, b)$ using only Theorem 3.
33. Let n be a positive integer with prime factorization $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$ where the p_i are distinct primes and $n_i \geq 1$ for each i .
 - Show that n has $(n_1 + 1)(n_2 + 1) \cdots (n_r + 1)$ distinct positive divisors.
 - Write down all the positive divisors of 340, 108, p^n , $p^2 q$, where p and q are distinct primes.

- (c) How many positive divisors does n have if $n = 25200$; $n = 41472$?
34. If $m \geq 1$ and $n \geq 1$ are relatively prime integers and nm is the square of an integer, show that both m and n are squares. Is this result true if m and n are not relatively prime?
35. If $\gcd(m, n) = 1$, where $m \geq 1$ and $n \geq 1$, and if $d | mn$, show that $d = m_1 n_1$ for some $m_1 | m$ and $n_1 | n$. [Hint: Theorem 7.]
36. Do Exercise 35 without assuming that $\gcd(m, n) = 1$. [Hint: If $0 \leq e \leq f + g$, where $f \geq 0$ and $g \geq 0$ are integers, show that e can be written $e = f_1 + g_1$, where $0 \leq f_1 \leq f$ and $0 \leq g_1 \leq g$. Use Theorem 8.]
37. Let $a \geq 1$ and $b \geq 1$ be integers. Show that there exist integers $u \geq 1$ and $v \geq 1$ such that $u | a$, $v | b$, $\gcd(u, v) = 1$, and $\text{lcm}(u, v) = ab$. [Hint: Theorem 9.]
38. If q is a rational number such that q^2 is an integer, show that q is an integer. [Hint: If $m^2 | n^2$, show that $m | n$ using Theorem 7.]
39. (a) Show that every prime $p > 2$ has the form $p = 4k + 1$ or $p = 4k + 3$.
 (b) Modify the proof of Theorem 10 to show that there are infinitely many primes of the form $4k + 3$.
40. A school has n lockers in a row along one side of a hall. The n students run down the hall one after the other. The first student closes all the lockers; then the second opens doors 2, 4, 6, ...; the third changes doors 3, 6, 9, ... (that is, opens a door if it is closed and closes it if it is open); the fourth student changes doors 4, 8, 12, ..., and so on. When all n students have gone through, which locker doors remain closed? Prove your answer. [Hint: Exercise 33(a).]
41. Compute the following:
 (a) $\gcd(28665, 22869)$ and $\text{lcm}(28665, 22869)$
 (b) $\gcd(231, 273, 429)$ and $\text{lcm}(231, 273, 429)$
 (c) $\gcd(1365, 1911, 1155, 1925)$ and $\text{lcm}(1365, 1911, 1155, 1925)$
42. Show that $\gcd(a, b, c) = \gcd[a, \gcd(b, c)]$.
43. Let $d = \gcd(a_1, a_2, a_3, \dots, a_k)$, where the a_i are positive integers. Show that integers x_1, x_2, \dots, x_k exist such that $d = x_1 a_1 + \dots + x_k a_k$. [Hint: Let m be the smallest member of $X = \{x_1 a_1 + \dots + x_k a_k \mid x_i \in \mathbb{Z}, x_1 a_1 + \dots + x_k a_k \geq 1\}$, and show that $m = d$. See the proof of Theorem 3.]
44. Let $b \geq 2$ be a fixed integer. If $n \geq 0$ is any integer, show that n can be written in the form $n = r_t b^t + r_{t-1} b^{t-1} + \dots + r_1 b + r_0$, where $t \geq 0$ and $0 \leq r_i < b$ for all i . Show further that these integers r_i and t are uniquely determined by n . This expression is called the **base b representation** of n .
45. Let $m \geq 1$ and $n \geq 1$ be integers.
 (a) If $m = qn + r$, $q, r \in \mathbb{Z}$, $0 \leq r < n$, show that $2^m - 1 = x(2^n - 1) + (2^r - 1)$ for some $x \in \mathbb{Z}$, where $0 \leq (2^r - 1) < 2^n - 1$.
 (b) If $d = \gcd(m, n)$, show that $\gcd(2^m - 1, 2^n - 1) = 2^d - 1$. [Hint: Get d by the Euclidean algorithm and use (a).]

1.3 INTEGERS MODULO n

Two integers a and b are said to have the same **parity** if both are even or both are odd, that is, if $2 | (a - b)$. The following definition extends this idea and introduces an important equivalence on the set \mathbb{Z} of integers. Let $n \geq 2$ be an integer.

Then integers a and b are said to be **congruent modulo n** if $n|(a - b)$. In this case we write $a \equiv b \pmod{n}$ and refer to n as the **modulus**.

Thus, we have $2 \equiv 5 \pmod{3}$, $21 \equiv 16 \pmod{5}$, and $-4 \equiv 2 \pmod{6}$. The expression $21832 \equiv 32 \pmod{100}$ explains why we can test whether an integer is divisible by 100 by looking at the last two digits. Note that $a \equiv 0 \pmod{n}$ if and only if $n | a$. We assume that $n \geq 2$ because congruence modulo 0 or 1 is of no interest (verify).

As the notation \equiv suggests, congruence modulo n is an equivalence relation on \mathbb{Z} .¹¹ The notation is justified in Theorem 1 and the proof is left as Exercise 6(a).

Theorem 1. Congruence modulo n is an equivalence on \mathbb{Z} ; that is:

- (1) $a \equiv a \pmod{n}$ for every integer a .
- (2) If $a \equiv b \pmod{n}$, then $b \equiv a \pmod{n}$.
- (3) If $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, then $a \equiv c \pmod{n}$.

If a is an integer, its equivalence class $[a]$ with respect to congruence modulo n is called its **residue class modulo n** , and we write $\bar{a} = [a]$ for convenience:

$$\bar{a} = [a] = \{x \in \mathbb{Z} \mid x \equiv a \pmod{n}\}.$$

The following result will be used frequently below.

Theorem 2. Given $n \geq 2$, $\bar{a} = \bar{b}$ if and only if $a \equiv b \pmod{n}$.

Proof. Suppose $\bar{a} = \bar{b}$. Since $a \in \bar{a}$, we have $a \in \bar{b}$, so $a \equiv b$. Conversely, let $a \equiv b$. Since \bar{a} and \bar{b} are sets, we must show that $\bar{a} \subseteq \bar{b}$ and $\bar{b} \subseteq \bar{a}$. If $x \in \bar{a}$, then $x \equiv a$; so, as $a \equiv b$, we have $x \equiv b$ by (3) of Theorem 1. This proves that $\bar{a} \subseteq \bar{b}$. Since $b \equiv a$ by (2) of Theorem 1, a similar proof shows that $\bar{b} \subseteq \bar{a}$. ■

Residue classes are easy to describe. For example, if $n = 2$,

$$\begin{aligned}\bar{0} &= \{x \in \mathbb{Z} \mid x \equiv 0 \pmod{2}\} = \text{the set of even integers} \\ \bar{1} &= \{x \in \mathbb{Z} \mid x \equiv 1 \pmod{2}\} = \text{the set of odd integers}\end{aligned}$$

In general, if a is an integer, the division algorithm gives $a = qn + r$, where $0 \leq r \leq n - 1$, so $a \equiv r \pmod{n}$. Thus every residue class modulo n appears in the list $\bar{0}, \bar{1}, \bar{2}, \dots, \bar{n-1}$. In fact it appears exactly once.

Theorem 3. Let $n \geq 2$ be an integer.

- (1) If $a \in \mathbb{Z}$, then $\bar{a} = \bar{r}$ for some r where $0 \leq r \leq n - 1$.
- (2) The residue classes $\bar{0}, \bar{1}, \bar{2}, \dots, \bar{n-1}$ modulo n are distinct.

Proof. It remains to verify (2). Suppose $\bar{r} = \bar{s}$, where $0 \leq r \leq n - 1$ and $0 \leq s \leq n - 1$. We may assume that $r \leq s$. Then $\bar{r} = \bar{s}$ means that $r \equiv s \pmod{n}$, so $s - r$ is an integral multiple of n such that $0 \leq s - r \leq n - 1$. This implies that $r = s$. ■

The set of all residue classes modulo n is denoted

$$\mathbb{Z}_n = \{\bar{0}, \bar{1}, \bar{2}, \dots, \bar{n-1}\}$$

¹¹See Section 0.4 for a discussion on equivalence relations.

and is called the set of **integers modulo n** . Thus, (2) of Theorem 3 is the assertion that $|\mathbb{Z}_n| = n$. In particular, $\mathbb{Z}_2 = \{\bar{0}, \bar{1}\}$, $\mathbb{Z}_3 = \{\bar{0}, \bar{1}, \bar{2}\}$, and so on.¹²

Example 1. Locate $\bar{48}$ and $\bar{-16}$ in $\mathbb{Z}_7 = \{\bar{0}, \bar{1}, \bar{2}, \bar{3}, \bar{4}, \bar{5}, \bar{6}\}$.

Solution. It seems that $\bar{48}$ does not appear. However, $48 \equiv 6 \pmod{7}$ means that $\bar{48} = \bar{6}$ does indeed occur. Similarly, $-16 \equiv 5 \pmod{7}$, so $\bar{-16} = \bar{5}$ also appears. \square

Example 2. If a is an odd integer, show that $\bar{a} = \bar{1}$ or $\bar{a} = \bar{3}$ in $\mathbb{Z}_4 = \{\bar{0}, \bar{1}, \bar{2}, \bar{3}\}$.

Solution. We know that \bar{a} is one of $\bar{0}, \bar{1}, \bar{2}$, or $\bar{3}$ in \mathbb{Z}_4 . If $\bar{a} = \bar{2}$, then $a \equiv 2 \pmod{4}$, so $a - 2 = 4q$ for some integer q . This means that a is even, contrary to assumption. So $\bar{a} \neq \bar{2}$ and, similarly, $\bar{a} \neq \bar{0}$. The only other possibilities are $\bar{a} = \bar{1}$ and $\bar{a} = \bar{3}$. \square

Example 3. In \mathbb{Z}_n , show that $\bar{a} = \bar{0}$ if and only if $n|a$.

Solution. By Theorem 2, $\bar{a} = \bar{0}$ means that $a \equiv 0 \pmod{n}$, that is, $n|a$. \square

Congruence modulo n is compatible with addition and multiplication of integers in the following sense. Let a, a_1, b , and b_1 denote integers.

$$\text{If } \begin{cases} a \equiv a_1 \pmod{n} \\ b \equiv b_1 \pmod{n} \end{cases} \text{ then } \begin{aligned} a + b &\equiv a_1 + b_1 \pmod{n} \\ ab &\equiv a_1 b_1 \pmod{n} \end{aligned} \quad (*)$$

In fact, let $a - a_1 = pn$ and $b - b_1 = qn$, where p and q are integers. Adding these equations gives $(a + b) - (a_1 + b_1) = (p + q)n$, and this implies that $a + b \equiv a_1 + b_1 \pmod{n}$. Similarly, multiplying the equations $a = a_1 + pn$ and $b = b_1 + qn$ gives $ab \equiv a_1 b_1 \pmod{n}$.

Condition (*) means that the arithmetic of \mathbb{Z} extends naturally to \mathbb{Z}_n as follows: We define addition and multiplication of residue classes \bar{a} and \bar{b} in \mathbb{Z}_n by

$$\bar{a} + \bar{b} = \overline{a + b} \quad \text{and} \quad \bar{a}\bar{b} = \overline{ab}. \quad (**)$$

Of course, we must verify that these operations are well defined, that is, we must check that they do not depend on which generators are used for the residue classes \bar{a} and \bar{b} . More precisely, suppose that

$$\bar{a} = \bar{a}_1 \quad \text{and} \quad \bar{b} = \bar{b}_1,$$

where $a \neq a_1$ and $b \neq b_1$ are possible. If we add these classes as \bar{a} and \bar{b} , (**) gives their sum as $\bar{a} + \bar{b}$, but if we represent the classes as \bar{a}_1 and \bar{b}_1 , their sum is $\bar{a}_1 + \bar{b}_1$. Clearly, the definition of addition makes no sense unless $\bar{a} + \bar{b} = \bar{a}_1 + \bar{b}_1$. But $a \equiv a_1$ and $b \equiv b_1$ by Theorem 2, so $a + a_1 \equiv b + b_1$ by (*), so $\bar{a} + \bar{b} = \bar{a}_1 + \bar{b}_1$, as required. Similarly, (*) shows that $\bar{a}\bar{b} = \bar{a}_1\bar{b}_1$, so the definition of multiplication also makes sense. In other words, addition and multiplication of residue classes are well defined by (**).

Example 4. In \mathbb{Z}_6 compute $\bar{3} + \bar{5}$ and $\bar{3} \cdot \bar{5}$.

Solution. The definition gives $\bar{3} + \bar{5} = \bar{8} = \bar{2}$, because $8 \equiv 2 \pmod{6}$. Similarly, $\bar{3} \cdot \bar{5} = \bar{15} = \bar{3}$. \square

¹²Note that \bar{a} means different things in \mathbb{Z}_2 , \mathbb{Z}_3 , ..., so to avoid ambiguity, perhaps we should denote residue classes \bar{a} in such a way that the modulus is apparent (say, ${}^2\bar{a}$ and ${}^3\bar{a}$). However, this is rarely done in practice as the modulus is usually clear from the context.

Theorem 4 collects several properties of these operations in \mathbb{Z}_n , each of which is the analogue of the corresponding property for \mathbb{Z} .

Theorem 4. Let $n \geq 2$ be a fixed modulus and let a, b , and c denote arbitrary integers. Then the following hold in \mathbb{Z}_n .

- (1) $\bar{a} + \bar{b} = \bar{b} + \bar{a}$ and $\bar{a}\bar{b} = \bar{b}\bar{a}$.
- (2) $\bar{a} + (\bar{b} + \bar{c}) = (\bar{a} + \bar{b}) + \bar{c}$ and $\bar{a}(\bar{b}\bar{c}) = (\bar{a}\bar{b})\bar{c}$.
- (3) $\bar{a} + \bar{0} = \bar{a}$ and $\bar{a}\bar{1} = \bar{a}$.
- (4) $\bar{a} + \overline{-\bar{a}} = \bar{0}$.
- (5) $\bar{a}(\bar{b} + \bar{c}) = \bar{a}\bar{b} + \bar{a}\bar{c}$.

Proof. We prove (5) and leave the rest as Exercise 6(b). Thus,

$$\begin{aligned} \bar{a}(\bar{b} + \bar{c}) &= \bar{a}(\overline{\bar{b} + \bar{c}}) \quad (\text{definition of addition in } \mathbb{Z}_n) \\ &= \overline{\bar{a}(\bar{b} + \bar{c})} \quad (\text{definition of multiplication in } \mathbb{Z}_n) \\ &= \overline{\bar{a}\bar{b} + \bar{a}\bar{c}} \quad (\text{property of } \mathbb{Z}) \\ &= \overline{\bar{a}\bar{b}} + \overline{\bar{a}\bar{c}} \quad (\text{definition of addition in } \mathbb{Z}_n) \\ &= \bar{a}\bar{b} + \bar{a}\bar{c} \quad (\text{definition of multiplication in } \mathbb{Z}_n), \end{aligned}$$

which proves (5). ■

These properties enable us to do arithmetic in \mathbb{Z}_n in much the same way as in \mathbb{Z} . In particular, (3) shows that $\bar{0}$ and $\bar{1}$ play roles in \mathbb{Z}_n analogous to those of 0 and 1 in \mathbb{Z} . For this reason, $\bar{0}$ and $\bar{1}$ are called the *zero* of \mathbb{Z}_n and the *unity* of \mathbb{Z}_n , respectively. Similarly, because of (4), $\overline{-\bar{a}}$ is called the *negative* of \bar{a} in \mathbb{Z}_n , and is denoted $\overline{-\bar{a}} = -\bar{a}$. Then *subtraction* in \mathbb{Z}_n is defined by

$$\bar{a} - \bar{b} = \bar{a} + \overline{-\bar{b}} = \overline{\bar{a} - \bar{b}},$$

an operation used much as it is in \mathbb{Z} .

Now consider the addition and multiplication tables for $\mathbb{Z}_6 = \{\bar{0}, \bar{1}, \bar{2}, \bar{3}, \bar{4}, \bar{5}\}$:

$+$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	\times	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$						
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{2}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{3}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{4}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{5}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$

These tables reveal many differences between the arithmetic of \mathbb{Z}_6 and that of \mathbb{Z} . For example, while 0 and 1 are the only integers k in \mathbb{Z} with the property that $k^2 = k$, each of $\bar{0}, \bar{1}, \bar{3}$, and $\bar{4}$ enjoy this property in \mathbb{Z}_6 . Another difference is that if $ab = ac$ in \mathbb{Z} and $a \neq 0$, then $b = c$. But $\bar{4} \cdot \bar{2} = \bar{4} \cdot \bar{5}$ in \mathbb{Z}_6 , and $\bar{4} \neq \bar{0}$, but $\bar{2} \neq \bar{5}$. Hence, we must be careful about “cancellation” in \mathbb{Z}_n . In fact, this concern is related to another difference between \mathbb{Z} and \mathbb{Z}_n . If $ab = 0$ in \mathbb{Z} , then $a = 0$ or $b = 0$. However, this need not hold in \mathbb{Z}_n . For example, $\bar{2} \cdot \bar{3} = \bar{0}$ in \mathbb{Z}_6 , but $\bar{2} \neq \bar{0}$ and $\bar{3} \neq \bar{0}$.

In Examples 5–7, we use the arithmetic of \mathbb{Z}_n to deduce facts about \mathbb{Z} . The connection is the fact (in Theorem 2) that $\bar{a} = \bar{b}$ in \mathbb{Z}_n means that $a \equiv b \pmod{n}$.

Example 5. Show that $a^5 \equiv a \pmod{5}$ holds for all integers a .

Solution. For an integer a , it suffices by Theorem 2 to show that $\bar{a}^5 = \bar{a}$ in \mathbb{Z}_5 . Because \bar{a} equals $\bar{0}, \bar{1}, \bar{2}, \bar{3}$, or $\bar{4}$, we examine each case separately.

- If $\bar{a} = \bar{0}$, then $\bar{a}^5 = \bar{0}^5 = \bar{0} = \bar{a}$.
- If $\bar{a} = \bar{1}$, then $\bar{a}^5 = \bar{1}^5 = \bar{1} = \bar{a}$.
- If $\bar{a} = \bar{2}$, then $\bar{a}^5 = \bar{2}^5 = \bar{2}^3 \cdot \bar{2}^2 = \bar{3} \cdot \bar{4} = \bar{2} = \bar{a}$.
- If $\bar{a} = \bar{3}$, then $\bar{a}^5 = \bar{3}^5 = \bar{9} \cdot \bar{27} = \bar{4} \cdot \bar{2} = \bar{3} = \bar{a}$.
- If $\bar{a} = \bar{4}$, then $\bar{a}^5 = \bar{4}^5 = \bar{16} \cdot \bar{64} = \bar{1} \cdot \bar{4} = \bar{4} = \bar{a}$.

Hence, $\bar{a}^5 = \bar{a}$ in every case, so $a^5 \equiv a \pmod{5}$ for all integers a . □

Example 5 is a special case of Fermat's theorem, which, for any prime p , asserts that $a^p \equiv a \pmod{p}$ for all integers a . We return to it later (Theorem 8).

Example 6. What is the remainder when 4^{119} is divided by 7?

Solution. If we can show that $4^{119} \equiv r \pmod{7}$, where $0 \leq r \leq 6$, then r is the desired remainder. We do the computation in \mathbb{Z}_7 . Note that, as $\bar{4}^2 = \bar{2}$ in \mathbb{Z}_7 , we have $\bar{4}^3 = \bar{8} = \bar{1}$. With this in mind, divide the exponent 119 by 3 to get $119 = 3 \cdot 39 + 2$. Then,

$$\bar{4}^{119} = \bar{4}^{3 \cdot 39 + 2} = (\bar{4}^3)^{39} \cdot \bar{4}^2 = \bar{1}^{39} \cdot \bar{2} = \bar{2}.$$

Hence, $4^{119} \equiv 2 \pmod{7}$, so the required remainder is 2. □

If a is an integer in decimal notation, it is common knowledge that a is divisible by 2 or 5 if and only if the same is true of its unit digit. Example 7 gives a similar test for divisibility by 9.

Example 7. Casting Out Nines. Show that a positive integer is divisible by 9 if and only if the sum of its digits is divisible by 9.

Solution. If $a = d_r d_{r-1} \dots d_1 d_0$ in decimal notation, where d_0, d_1, \dots, d_r are the digits, then $a = d_0 + 10d_1 + 10^2d_2 + \dots + 10^r d_r$. Now $\bar{10} = \bar{1}$ in \mathbb{Z}_9 , so $\bar{10}^k = \bar{1}^k = \bar{1}$ for each k . Hence, in \mathbb{Z}_9 ,

$$\bar{a} = \bar{d}_0 + \bar{1} \cdot \bar{d}_1 + \bar{1}^2 \cdot \bar{d}_2 + \dots + \bar{1}^r \cdot \bar{d}_r = \overline{d_0 + d_1 + \dots + d_r}.$$

Thus, $a \equiv d_0 + d_1 + \dots + d_r \pmod{9}$, and the result follows from Example 3. □

These three examples show that the properties in Theorem 4 allow many of the operations of ordinary arithmetic to be carried out in \mathbb{Z}_n . However, these properties tell us nothing about how to solve an equation such as $\bar{a}x = \bar{b}$ in \mathbb{Z}_n . For example, consider

$$\bar{5}x = \bar{2}$$

in \mathbb{Z}_{17} . The desired solution (if there is one) is a residue class x in \mathbb{Z}_{17} , so x is one of $\bar{0}, \bar{1}, \bar{2}, \dots, \bar{16}$. Hence, one method is simply to try all these classes! If we do so, we find that $x = \bar{14}$ is the only solution. However, this method is impractical if the modulus is large.

A better approach is as follows. Suppose that a residue class \bar{b} can be found such that $\bar{b} \cdot \bar{5} = \bar{1}$. Then if we multiply both sides of the equation $\bar{5}x = \bar{2}$ by \bar{b} , the

result is $\bar{b} \cdot \bar{5}x = \bar{b} \cdot \bar{2}$, that is, $x = \bar{2}\bar{b}$. The class \bar{b} (if it exists) can again be found by trial and error. In fact $\bar{b} = \bar{7}$ works, so $x = \bar{2}\bar{7} = \bar{14}$, as before.

Fortunately, there is a systematic way of finding \bar{b} in \mathbb{Z}_{17} such that $\bar{b} \cdot \bar{5} = \bar{1}$. Note that 5 and 17 are relatively prime, so the euclidean algorithm can be used to express $\gcd(5, 17) = 1$ as a linear combination of 5 and 17. In fact, we have

$$17 = 3 \cdot 5 + 2 \quad \text{and then} \quad 5 = 2 \cdot 2 + 1;$$

so, eliminating remainders, $1 = 5 - 2(17 - 3 \cdot 5) = 7 \cdot 5 - 2 \cdot 17$. This implies that $7 \cdot 5 \equiv 1 \pmod{17}$, and so $\bar{7} \cdot \bar{5} = \bar{1}$ in \mathbb{Z}_{17} . This gives $\bar{b} = \bar{7}$.

This method clearly generalizes. For a modulus $n \geq 2$ and an integer a , a residue class \bar{b} in \mathbb{Z}_n is called an **inverse** of \bar{a} if $\bar{b}\bar{a} = \bar{1}$ in \mathbb{Z}_n . If \bar{a} has an inverse, that inverse is unique (Exercise 23) and we say \bar{a} is **invertible**. Theorem 5 characterizes when an inverse exists, and the proof shows that (as above) the euclidean algorithm can be used to find it.

Theorem 5. Let a and n be integers with $n \geq 2$. Then \bar{a} has an inverse in \mathbb{Z}_n if and only if a and n are relatively prime.

Proof. If a and n are relatively prime, then $1 = \gcd(a, n)$ is a linear combination of a and n (by Theorem 4 §1.2), say $1 = ba + cn$, where b and c are integers. Hence, $ba \equiv 1 \pmod{n}$, so $\bar{b}\bar{a} = \bar{1}$ by Theorem 2. Conversely, if b exists such that $\bar{b}\bar{a} = \bar{1}$, then $ba \equiv 1 \pmod{n}$. Thus, $n|(1 - ba)$, say $1 - ba = qn$ for some integer q . But then $1 = ba + qn$, so a and n are relatively prime (again by Theorem 4 §1.2). ■

Example 8. Find the inverse of $\bar{16}$ in \mathbb{Z}_{35} and use it to solve $\bar{16}x = \bar{9}$ in \mathbb{Z}_{35} .

Solution. The inverse exists as $\gcd(35, 16) = 1$. The euclidean algorithm gives

$$35 = 2 \cdot 16 + 3 \quad \text{and then} \quad 16 = 5 \cdot 3 + 1,$$

so $1 = 16 - 5(35 - 2 \cdot 16) = 11 \cdot 16 - 5 \cdot 35$. Thus, $11 \cdot 16 \equiv 1 \pmod{35}$, and so $\bar{11}$ is the inverse of $\bar{16}$ in \mathbb{Z}_{35} . Now multiply the equation $\bar{16}x = \bar{9}$ by $\bar{11}$ to obtain $\bar{11} \cdot \bar{16}x = \bar{11} \cdot \bar{9}$; that is, $x = \bar{99} = \bar{29}$. □

Example 9. Find the elements in \mathbb{Z}_9 that have inverses.

Solution. The members of \mathbb{Z}_9 are of the form \bar{r} , where $r = 0, 1, 2, \dots, 8$. Since $9 = 3^2$, r is relatively prime to 9 if and only if r is not a multiple of 3. Hence, $\bar{1}, \bar{2}, \bar{4}, \bar{5}, \bar{7}$, and $\bar{8}$ will all have inverses. Indeed, $\bar{1}$ and $\bar{8}$ are both self-inverse, whereas $\bar{2}$ and $\bar{5}$ are inverses of each other as are $\bar{4}$ and $\bar{7}$. □

Example 10. Solve the system $\begin{cases} \bar{5}x + \bar{8}y = \bar{2} \\ \bar{3}x + \bar{2}y = \bar{1} \end{cases}$ of equations in \mathbb{Z}_{11} .

Solution. The usual techniques apply. Since $\bar{4} \cdot \bar{3} = \bar{1}$, we eliminate y by first multiplying the second equation by $\bar{4}$ to get $x + \bar{8}y = \bar{4}$. Subtract this from the first equation to get $\bar{4}x = -\bar{2} = \bar{9}$. Now $\bar{3}$ is the inverse of $\bar{4}$ in \mathbb{Z}_{11} , so multiplication by $\bar{3}$ gives $x = \bar{3} \cdot \bar{9} = \bar{5}$. Then the last equation gives $\bar{2}y = \bar{1} - \bar{3}x = \bar{8}$. Finally, $\bar{6}$ is the inverse of $\bar{2}$, so $y = \bar{6} \cdot \bar{8} = \bar{4}$. □

If a is a real number, an expression $x^2 + ax$ becomes a square if $(\frac{1}{2}a)^2$ is added: $x^2 + ax + (\frac{1}{2}a)^2 = (x + \frac{1}{2}a)^2$. This process is called **completing the square**, and it works in \mathbb{Z}_n provided $\bar{2}$ has an inverse in \mathbb{Z}_n (that is, if n is odd).

Example 11. Solve the quadratic $x^2 + \bar{3}x + \bar{9} = \bar{0}$ in \mathbb{Z}_{13} .

Solution. First subtract $\bar{9}$ from both sides to obtain $x^2 + \bar{3}x = -\bar{9} = \bar{4}$. The inverse of $\bar{2}$ in \mathbb{Z}_{13} is $\bar{7}$, so we complete the square on the left by adding $(\bar{7} \cdot \bar{3})^2 = \bar{8}^2 = \bar{12}$ to both sides. The result is $x^2 + \bar{3}x + \bar{12} = \bar{4} + \bar{12}$, that is, $(x + \bar{8})^2 = \bar{3}$. Now \mathbb{Z}_{13} has 13 elements and, by inspection, only 2 of them square to $\bar{3}$, namely, $\bar{4}$ and $-\bar{4} = \bar{9}$. Hence, $x + \bar{8} = \bar{4}$ or $x + \bar{8} = \bar{9}$, and so $x = \bar{9}$ and $x = \bar{1}$ are the solutions. \square

Note that there are *two* solutions in Example 11. The reason is that $\bar{3}$ has two “square roots” in \mathbb{Z}_{13} : $\bar{4}$ and $-\bar{4} = \bar{9}$. However, other situations are possible: In \mathbb{Z}_7 , $\bar{3}$ has no square root, whereas in \mathbb{Z}_{27} , $\bar{9}$ has six square roots, $\bar{3}$ and $-\bar{3} = \bar{24}$, $\bar{6}$ and $-\bar{6} = \bar{21}$, and finally $\bar{12}$ and $-\bar{12} = \bar{15}$.

The following fact about congruences is useful in number theory and computer science, and was known to the Chinese in the fourth century.

Theorem 6. Chinese Remainder Theorem. Let m and n be relatively prime integers. If s and t are arbitrary integers, there exists a solution $x \in \mathbb{Z}$ to the simultaneous congruences

$$x \equiv s \pmod{m} \quad \text{and} \quad x \equiv t \pmod{n}.$$

Proof. Since $\gcd(m, n) = 1$, the Euclidean algorithm gives p and q in \mathbb{Z} such that $1 = mp + nq$. Take

$$x = (mp)t + (nq)s.$$

Then $x - s = mpt + (nq - 1)s = mp(t - s)$, so $x \equiv s \pmod{m}$. A similar argument gives $x \equiv t \pmod{n}$. \blacksquare

The nice thing about Theorem 6 is that the proof gives an algorithm for finding the solution x : The Euclidean algorithm gives p and q such that $1 = mp + nq$, and the solution is $x = mpt + nqs$. Furthermore, this method can be iterated to solve a system of more than two congruences, provided that only the moduli are relatively prime in pairs. To illustrate, let m_1, m_2 , and m_3 be integers relatively prime in pairs. Given arbitrary integers s_1, s_2 , and s_3 , we want to find an integer x such that

$$x \equiv s_i \pmod{m_i} \quad \text{for each } i = 1, 2, 3.$$

The Chinese remainder theorem yields a such that $a \equiv s_i \pmod{m_i}$ for $i = 1, 2$. Since $m_1 m_2$ and m_3 are relatively prime, apply the Chinese remainder theorem again to obtain x such that

$$x \equiv a \pmod{m_1 m_2} \quad \text{and} \quad x \equiv s_3 \pmod{m_3}.$$

But then $x \equiv a \pmod{m_1}$, so since $a \equiv s_1 \pmod{m_1}$, we have $x \equiv s_1 \pmod{m_1}$. Similarly, $x \equiv s_2 \pmod{m_2}$.

In general, if m_1, m_2, \dots, m_k are relatively prime in pairs, and if s_1, s_2, \dots, s_k are arbitrary integers, then there exists $x \in \mathbb{Z}$ such that

$$x \equiv s_i \pmod{m_i} \quad \text{for each } i = 1, 2, \dots, k.$$

These general systems of congruences are important in computer science because they provide a method for doing arithmetic with integers that exceed the *word size* of the computer (the largest integer that can be used in machine arithmetic).

The only elements of \mathbb{Z} that have an inverse in \mathbb{Z} are 1 and -1 (because $\frac{1}{k}$ does not lie in \mathbb{Z} if $k \neq 1, -1$). Thus, \mathbb{Z} resembles \mathbb{Z}_6 in this respect (see the table following Theorem 4). At the other extreme, *every* nonzero real number $x \neq 0$ has an inverse $\frac{1}{x}$ in \mathbb{R} . Theorem 7 characterizes when this happens in \mathbb{Z}_n .

Theorem 7. *The following are equivalent for an integer $n \geq 2$.*

- (1) *Every element $\bar{a} \neq \bar{0}$ in \mathbb{Z}_n has an inverse.*
- (2) *If $\bar{a}\bar{b} = \bar{0}$ in \mathbb{Z}_n , then either $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$.*
- (3) *n is a prime.*

Proof. We prove that (1) \Rightarrow (2), (2) \Rightarrow (3), and (3) \Rightarrow (1).

(1) \Rightarrow (2). Assume (1) is true and let $\bar{a}\bar{b} = \bar{0}$ in \mathbb{Z}_n . If $\bar{a} = \bar{0}$, there is nothing to prove. Otherwise, \bar{a} has an inverse by (1), say $\bar{c}\bar{a} = \bar{1}$. Then we multiply both sides of $\bar{a}\bar{b} = \bar{0}$ by \bar{c} to get $\bar{c}\bar{a}\bar{b} = \bar{c}\bar{0}$; that is, $\bar{b} = \bar{0}$.

(2) \Rightarrow (3). If n is not prime, let $n = ab$, where $2 \leq a < n$ and $2 \leq b < n$. But then $\bar{a}\bar{b} = \bar{n} = \bar{0}$, where $\bar{a} \neq \bar{0}$ and $\bar{b} \neq \bar{0}$. This contradicts (2), so the assumption that n is not prime cannot be valid.

(3) \Rightarrow (1). If n is prime, let $\bar{a} \neq \bar{0}$ in \mathbb{Z}_n . Then $\gcd(a, n) = 1$ (because otherwise $\gcd(a, n) = n$, so $n|a$). But then $1 = ba + cn$ for integers b and c (by Theorem 4 §1.2), so $ba \equiv 1 \pmod{n}$. Thus, $\bar{b}\bar{a} = \bar{1}$ in \mathbb{Z}_n , proving (1). ■

Hence, if p is a prime, \mathbb{Z}_p has the property that every nonzero element has an inverse. This is also true of the real numbers \mathbb{R} , and such systems are called **fields**.

The following consequence of Theorem 7 will be referred to later.

Corollary. Wilson's Theorem. *If p is a prime, then $(p-1)! \equiv -1 \pmod{p}$.*

Proof. We write $\bar{a} = a$ in \mathbb{Z}_p for convenience. Since p is prime, each element $1, 2, 3, \dots, p-1$ in \mathbb{Z}_p has an inverse by Theorem 7. Hence, pairs of inverses in the product $(p-1)! = 1 \ 2 \ 3 \cdots (p-1)$ will cancel leaving only the self-inverse elements 1 and -1 (Exercise 26). Thus, $(p-1)! = 1(-1) = -1$ in \mathbb{Z}_p , as required. ■

Example 12. Write down the multiplication table of \mathbb{Z}_5 and illustrate Theorem 7.

Solution. The first row and column of the table consist entirely of zeros (true for any modulus), but the fact that no other entry equals $\bar{0}$ verifies (2) of Theorem 7. Similarly, the fact that every row (or column) except the first contains $\bar{1}$ verifies (1) of Theorem 7.

\times	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{4}$	$\bar{1}$	$\bar{3}$
$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{1}$	$\bar{4}$	$\bar{2}$
$\bar{4}$	$\bar{0}$	$\bar{4}$	$\bar{3}$	$\bar{2}$	$\bar{1}$

The simplest situation in which Theorem 7 applies is when $n = 2$. In this case, $\mathbb{Z}_2 = \{\bar{0}, \bar{1}\}$ and the addition and multiplication tables are as follows:

$+$	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{1}$
$\bar{1}$	$\bar{1}$	$\bar{0}$

\times	$\bar{0}$	$\bar{1}$
$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$

This is binary arithmetic, which is important in the design of computers.

We conclude with a famous theorem of Pierre de Fermat. In Example 5, we showed that $a^5 \equiv a \pmod{5}$ holds for all integers a . In fact, it holds if we replace 5 by any prime.

Theorem 8. Fermat's Theorem. *If p is a prime, then*

$$a^p \equiv a \pmod{p} \quad \text{for all integers } a.$$

In fact, $a^{p-1} \equiv 1 \pmod{p}$ for all integers a that are relatively prime to p .

Proof. We must show that $\bar{a}^p = \bar{a}$ in \mathbb{Z}_p . Because this equation is true if $\bar{a} = \bar{0}$, it suffices to show that $\bar{a}^{p-1} = \bar{1}$ in \mathbb{Z}_p whenever $\bar{a} \neq \bar{0}$. But if $\bar{a} \neq \bar{0}$, then \bar{a} has an inverse in \mathbb{Z}_p by Theorem 7, say $\bar{b}\bar{a} = \bar{1}$. Now multiply all the nonzero elements in \mathbb{Z}_p by \bar{a} to obtain

$$\bar{a}\bar{1}, \bar{a}\bar{2}, \dots, \bar{a}(\bar{p}-\bar{1}).$$

These are all distinct (because $\bar{a}\bar{r} = \bar{a}\bar{s}$ yields $\bar{r} = \bar{s}$ after multiplication by \bar{b}) and none equals $\bar{0}$, so they must be the set of *all* nonzero elements $\bar{1}, \bar{2}, \dots, \bar{p}-\bar{1}$ in some order. In particular, the products are the same, and we obtain

$$\bar{a}^{p-1}(\bar{1}\bar{2}\cdots\bar{p}-\bar{1}) = \bar{1}\bar{2}\cdots\bar{p}-\bar{1}.$$

But the element $\bar{1}\bar{2}\cdots\bar{p}-\bar{1}$ is invertible in \mathbb{Z}_p (Exercise 24). Hence, multiplication by its inverse gives $\bar{a}^{p-1} = \bar{1}$, which is what we wanted. ■

Note that Fermat's theorem fails if p is not prime; for example, $2^4 \not\equiv 2 \pmod{4}$.

Fermat's theorem is important in number theory, and the following result will be referred to several times. To state it, we use the following useful observation (Exercise 36): If prime $p > 2$ is a prime, then $p \equiv 1 \pmod{4}$ or $p \equiv 3 \pmod{4}$.

Corollary. *Let $p > 2$ be a prime.*

- (1) *If $p \equiv 1 \pmod{4}$, then $x^2 = -1$ in \mathbb{Z}_p , where $x = \bar{1}\bar{2}\cdots\frac{1}{2}(p-1)$.*
- (2) *If $p \equiv 3 \pmod{4}$, then the equation $x^2 = -1$ has no solution in \mathbb{Z}_p .*

Proof. Write $\bar{a} = a$ in \mathbb{Z}_p for convenience.

- (1) We have $(p-1)! = -1$ by the Corollary to Theorem 7. Write

$$q = \frac{1}{2}(p+1)\cdots(p-2)(p-1).$$

Then,

$$xq = [1\bar{2}\cdots\frac{1}{2}(p-1)] [\frac{1}{2}(p+1)\cdots(p-2)(p-1)] = (p-1)! = -1.$$

Thus, it suffices to show that $q = x$. Now observe that we can write q as follows:

$$q = (-\frac{1}{2}(p-1))\cdots(-2)(-1).$$

Since $p \equiv 1 \pmod{4}$, the integer $\frac{1}{2}(p-1)$ is even. Hence, q has an even number of factors, and it follows that $q = x$ after all. This proves (1).

- (2) Let $p = 4n + 3$ in \mathbb{Z} . Suppose $a \in \mathbb{Z}_p$ satisfies $a^2 = -1$ in \mathbb{Z}_p ; we look for a contradiction. Since $a^{p-1} = 1$ by Fermat's theorem, we have

$$1 = a^{p-1} = a^{4n+2} = (a^2)^{2n+1} = (-1)^{2n+1} = -1 \text{ in } \mathbb{Z}_p,$$

a contradiction because $p > 2$. So $x^2 = -1$ has no solution in \mathbb{Z}_p , proving (2). ■

Clearly, a residue class \bar{a} is not the same thing as the integer a . However, because of the definitions $\bar{a} + \bar{b} = a + b$ and $\bar{a}\bar{b} = ab$ in \mathbb{Z}_n , the arithmetic of \mathbb{Z}_n closely resembles that of \mathbb{Z} —so much so that in subsequent chapters we adopt the following convention (used above in the Corollaries to Theorems 7 and 8):

Notational Convention. When working in \mathbb{Z}_n we frequently write the residue class \bar{a} simply as a .

Then $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$, and equations such as $3 \cdot 4 = 2$ and $2 + 3 = 0$ appear. This notation is harmless, once everyone knows that we are using it, and it facilitates hand calculations (the reader as probably been using it already!). Of course, when the convention causes confusion, we revert to the more formal \bar{a} notation.

Pierre De Fermat (1601–1685) Fermat was a lawyer by profession and served in the parliament in Toulouse, France. His mathematical work was a pastime, and he has been called “the prince of amateurs.” This appellation should not be taken as diminishing his stature, because he did first-rate work in several areas. He invented analytic geometry prior to Descartes and made contributions to the development of calculus. Along with Pascal, he is credited with starting the theory of probability.

However, he is most remembered for his work in number theory. Theorem 8 first appeared in a letter in 1640, and a proof was first published much later by Euler. Fermat published virtually nothing, and his results became known through letters to his friends (many to Mersenne) and as notes jotted in the margin of his copy of *Arithmetica* by Diophantus, usually with no proof. The most famous of these notes is the assertion that, if $n \geq 3$, positive integers x , y , and z do not exist such that $x^n + y^n = z^n$. This assertion has become known as “Fermat’s Last Theorem”, and he wrote that “I have found a truly remarkable proof but the margin was too small to contain it.” His intuition was so good that every other theorem that he claimed he could prove has been subsequently verified. However, despite the best efforts of the greatest mathematicians, the “Last Theorem” remained open for 300 years. But in 1997, in a spectacular display of mathematical virtuosity, Andrew Wiles of Princeton University finally proved the result. Wiles related Fermat’s conjecture to a problem in geometry, which he solved.

Exercises 1.3

1. In each case determine whether the statement is true or false.

(a) $40 \equiv 13 \pmod{9}$	(b) $-29 \equiv 1 \pmod{7}$
(c) $-29 \equiv 6 \pmod{7}$	(d) $132 \equiv 0 \pmod{11}$
(e) $8 \equiv 8 \pmod{n}$	(f) $3^4 \equiv 1 \pmod{5}$
(g) $8^4 \equiv 2 \pmod{13}$	
2. In each case find all integers k making the statement true.

(a) $4 \equiv 2k \pmod{7}$	(b) $12 \equiv 3k \pmod{10}$
(c) $3k \equiv k \pmod{9}$	(d) $5k \equiv k \pmod{15}$
3. Find all integers $k \geq 2$ such that

(a) $-3 \equiv 7 \pmod{k}$	(b) $7 \equiv -5 \pmod{k}$
(c) $3 \equiv k^2 \pmod{k}$	(d) $5 \equiv k \pmod{k^2}$
4. Find all integers $k \geq 2$ such that $k^2 \equiv 5k \pmod{15}$.
5. (a) Show that congruence modulo 0 is equality.
 (b) What can you say about congruence modulo 1?
6. (a) Prove Theorem 1.
 (b) Prove (1)–(4) of Theorem 4.

28. Find $x \in \mathbb{Z}$ such that $x \equiv 8 \pmod{10}$, $x \equiv 3 \pmod{9}$, and $x \equiv 2 \pmod{7}$.
29. (a) If $\bar{a}\bar{b} = \bar{0}$ in \mathbb{Z}_n and $\gcd(a, n) = 1$, show that $\bar{b} = \bar{0}$.
 (b) Show that \bar{a} is invertible in \mathbb{Z}_n if and only if $\bar{a}\bar{b} = \bar{0}$ implies that $\bar{b} = \bar{0}$.
30. Show that the following conditions on an integer $n \geq 2$ are equivalent.
 (1) $\bar{a}^2 = \bar{0}$ in \mathbb{Z}_n implies that $\bar{a} = \bar{0}$.
 (2) n is square free (that is, a product of distinct primes).
 [Hint: Theorem 5 §1.2.]
31. Show that the following conditions on an integer $n \geq 2$ are equivalent.
 (1) If \bar{a} is in \mathbb{Z}_n , then either \bar{a} is invertible or $\bar{a}^k = \bar{0}$ for some $k \geq 1$.
 (2) n is a power of a prime.
32. If $p \geq 3$ is a prime, show that every element of \mathbb{Z}_p has a $(p-2)$ th root. [Hint: Use Fermat's theorem to show that $f: \mathbb{Z}_p \rightarrow \mathbb{Z}_p$ is one-to-one, where $f(\bar{a}) = \bar{a}^{p-2}$. Apply Theorem 2 §0.3.]
33. Show that $2^{37} - 1$ is divisible by 223 and that $2^{32} + 1$ is divisible by 641. (Remarkably, $\frac{1}{223}(2^{37} - 1)$ is also prime.) Note: If p is a prime, numbers of the form $2^p - 1$ and $2^{2^n} + 1$ are called **Mersenne numbers** and **Fermat numbers**, respectively, and were once thought to be all primes.
34. Let a and n denote integers with $n \geq 2$, and write $d = \gcd(a, n)$.
 (a) Show that $ax \equiv b \pmod{n}$ has a solution if and only if $d|b$.
 (b) If $d = ra + sn$, r and s integers, show that $x_0 = r(b/d)$ is one solution.
 (c) If x_0 is any solution, show that there are exactly d solutions that are distinct modulo n : $\{x_0, x_0 + \frac{n}{d}, x_0 + 2\frac{n}{d}, \dots, x_0 + (d-1)\frac{n}{d}\}$. [Hint: If $ax \equiv b \pmod{n}$, show that $a(x - x_0) \equiv 0 \pmod{n}$, so $(a/d)(x - x_0) \equiv 0 \pmod{n/d}$ by Exercise 11 §1.2. Conclude that $x - x_0 \equiv 0 \pmod{n/d}$.]
 (d) Find all solutions to $15x \equiv 25 \pmod{35}$.
 (e) Find all solutions to $21x \equiv 14 \pmod{35}$.
 (f) Find all solutions to $21x \equiv 8 \pmod{33}$.
35. Let p be a prime. If $x^2 = \bar{1}$ in \mathbb{Z}_p , show that $x = \bar{1}$ or $x = -\bar{1}$.
36. Let p be a prime, show that either $p \equiv 1 \pmod{4}$ or $p \equiv 3 \pmod{4}$.
37. (a) Show that if $a^n \equiv a \pmod{n}$ holds for all integers a , the modulus n must be square free, that is, a product of distinct primes.
 (b) Show that $a^{561} \equiv a \pmod{561}$ for all integers a . [Hint: Use Theorem 5 §1.2 to reduce the problem to showing that $a^{561} \equiv a \pmod{p}$, where $p = 3, 11$, or 17 . In each case, use Fermat's theorem in the form $a^{p-1} \equiv 1 \pmod{p}$ whenever p does not divide a .]

1.4 PERMUTATIONS

A permutation of the numbers 1, 2, and 3 is a rearrangement of these numbers in a definite order. Thus, the six possibilities are

$$\begin{array}{ccccccc} 1 & 2 & 3 & 1 & 3 & 2 & 2 & 1 & 3 & 2 & 3 & 1 & 3 & 1 & 2 & 3 & 2 & 1 \end{array}$$

They can also be described as mappings $\{1, 2, 3\} \rightarrow \{1, 2, 3\}$:

$$\begin{array}{llllll} 1 \rightarrow 1 & 1 \rightarrow 1 & 1 \rightarrow 2 & 1 \rightarrow 2 & 1 \rightarrow 3 & 1 \rightarrow 3 \\ 2 \rightarrow 2 & 2 \rightarrow 3 & 2 \rightarrow 1 & 2 \rightarrow 3 & 2 \rightarrow 1 & 2 \rightarrow 2 \\ 3 \rightarrow 3 & 3 \rightarrow 2 & 3 \rightarrow 3 & 3 \rightarrow 1 & 3 \rightarrow 2 & 3 \rightarrow 1 \end{array}$$

We use this terminology of mappings to describe permutations.

If X and Y are sets, recall that a mapping $\alpha : X \rightarrow Y$ is a rule that assigns to every element x of X exactly one element $\alpha(x)$ of Y , called the image of x under α . Hence, the diagram

$$\begin{aligned} 1 &\rightarrow 1 \\ 2 &\rightarrow 3 \\ 3 &\rightarrow 2 \end{aligned}$$

describes the mapping $\alpha : \{1, 2, 3\} \rightarrow \{1, 2, 3\}$ given by the rule $\alpha(1) = 1$, $\alpha(2) = 3$, $\alpha(3) = 2$.

Now consider a mapping $\alpha : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, n\}$. Because such mappings occur frequently, we write $\alpha(k) = \alpha k$ for simplicity. Our interest is in when the images $\alpha 1, \alpha 2, \dots, \alpha n$ are a *permutation* of the numbers $1, 2, \dots, n$; that is, each element of $\{1, 2, \dots, n\}$ occurs *exactly once* in the list $\alpha 1, \alpha 2, \dots, \alpha n$. In other words, the function α is both one-to-one and onto (a **bijection**).¹³

Given an integer $n \geq 1$, write $X_n = \{1, 2, \dots, n\}$.

A **permutation** of X_n is a bijection $\sigma : X_n \rightarrow X_n$.

We call the set S_n of all permutations of X_n the **symmetric group of degree n** . Two permutations σ and τ in S_n are **equal** if they are equal as functions, that is, if $\sigma k = \tau k$ for all k in X_n .

To simplify the manipulation of these permutations, a matrix-type notation is useful. For example, if the permutation $\sigma : X_4 \rightarrow X_4$ is defined by $\sigma 1 = 3$, $\sigma 2 = 1$, $\sigma 3 = 4$, and $\sigma 4 = 2$, we write it as

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}.$$

Here the image of each element of $X_4 = \{1, 2, 3, 4\}$ is written below that element. In general, a permutation $\sigma \in S_n$ is written in matrix form as

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma 1 & \sigma 2 & \cdots & \sigma n \end{pmatrix}.$$

Hence, a typical member of S_n takes this form, where $\sigma 1, \sigma 2, \dots, \sigma n$ is the list of numbers $1, 2, \dots, n$ in a (possibly) different order.

Example 1. List the elements of S_3 in matrix notation.

Solution. There are six different permutations:

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

In general, to construct a permutation

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma 1 & \sigma 2 & \cdots & \sigma n \end{pmatrix},$$

we must choose the numbers $\sigma 1, \sigma 2, \dots, \sigma n$ from X_n so that they are all distinct. Hence, we have n choices for $\sigma 1$, then $n - 1$ choices for $\sigma 2$, then $n - 2$ choices for

¹³A review of one-to-one and onto mappings can be found in Section 0.3.

$\sigma 3$, and so on. Thus, σ can be chosen in $n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1 = n!$ ways, which proves the following theorem:

Theorem 1. *The set S_n of permutations of X_n has $|S_n| = n!$ elements.*

Let σ and τ be permutations in S_n . Both are mappings from X_n to X_n , and we write them as follows:

$$X_n \xrightarrow{\tau} X_n \xrightarrow{\sigma} X_n.$$

We then define the *composite* $\sigma\tau: X_n \rightarrow X_n$ by first applying τ and then σ :

$$(\sigma\tau)k = \sigma(\tau k), \quad \text{for all } k \in X_n.$$

Because both σ and τ are one-to-one and onto, these properties hold for the composite $\sigma\tau$ (see Theorem 3 §0.3). Hence, $\sigma\tau$ is again a permutation in S_n .

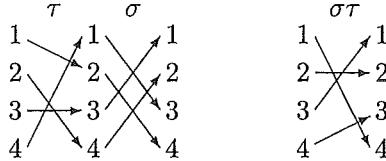
Example 2. Compute $\sigma\tau$ if

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \quad \text{and} \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 3 & 1 \end{pmatrix}.$$

Solution. Consider the action of $\sigma\tau$ on 1: $(\sigma\tau)1 = \sigma 2 = 4$. We can compute it directly from the matrix forms:

$$\sigma\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix}.$$

It is important to remember that, in computing $\sigma\tau$, we apply τ first and then σ . Thus, we read $1 \xrightarrow{\tau} 2$ from the matrix for τ , then $2 \xrightarrow{\sigma} 4$ from the matrix for σ . The result is $1 \xrightarrow{\sigma\tau} 4$, as indicated. Similarly, $2 \xrightarrow{\tau} 4 \xrightarrow{\sigma} 2$ leads to $2 \xrightarrow{\sigma\tau} 2$. We can read the entire action of $\sigma\tau$ in this manner. The following diagrams illustrate what is happening:



The action of $\sigma\tau$ is read from the first diagram by following the arrows. □

Note that $\sigma\tau \neq \tau\sigma$ in general: If σ and τ are as in Example 2,

$$\tau\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \end{pmatrix}$$

is not the same as $\sigma\tau$ (computed in Example 2). If it happens that $\sigma\tau = \tau\sigma$, we say that σ and τ *commute*. Thus, two permutations need not commute (but see Theorem 3). On the other hand, if σ, τ , and μ are three permutations in S_n then we always have

$$(\sigma\tau)\mu = \sigma(\tau\mu),$$

which we can easily verify directly (see Theorem 3 §0.3).

The **identity permutation** ε in S_n is defined as

$$\varepsilon = \begin{pmatrix} 1 & 2 & \cdots & n \\ 1 & 2 & \cdots & n \end{pmatrix}.$$

In other words, $\varepsilon k = k$ holds for every $k \in X_n$. It is easy to verify that

$$\varepsilon\sigma = \sigma = \sigma\varepsilon$$

holds for all $\sigma \in S_n$, so ε plays the role in S_n that 1 plays for multiplication of numbers.

Consider the permutation

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

in S_4 . The action of σ is obtained by *reading down*: $\sigma 1 = 3$, $\sigma 2 = 4$, $\sigma 3 = 2$, and $\sigma 4 = 1$. There is clearly another permutation in S_4 obtained by *reading up* $3 \rightarrow 1$, $4 \rightarrow 2$, $2 \rightarrow 3$, and $1 \rightarrow 4$. This new permutation is determined uniquely by σ ; In fact, it is the inverse of σ (denoted σ^{-1} as in Section 0.3). Thus,

$$\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix}.$$

In general, if $\sigma \in S_n$, the fact that $\sigma: X_n \rightarrow X_n$ is one-to-one and onto implies (Theorem 6 §0.3) that a uniquely determined permutation $\sigma^{-1}: X_n \rightarrow X_n$ exists (called the **inverse** of σ), which satisfies

$$\sigma(\sigma^{-1}k) = k \quad \text{and} \quad \sigma^{-1}(\sigma k) = k, \quad \text{for all } k \in X_n. \tag{*}$$

Equations (*) imply that each of σ and σ^{-1} reverses the action of the other and hence that we can indeed obtain the action of σ^{-1} from

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma 1 & \sigma 2 & \cdots & \sigma n \end{pmatrix}$$

by reading up.

Example 3. Find the inverse of $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 1 & 8 & 3 & 2 & 5 & 6 & 7 \end{pmatrix}$ in S_8 .

Solution. Reversing the action of σ gives $\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 5 & 4 & 1 & 6 & 7 & 8 & 3 \end{pmatrix}$. \square

If $\sigma \in S_n$, it is related to σ^{-1} by composition. Indeed, because the identity permutation ε in S_n satisfies $\varepsilon k = k$ for all $k \in X_n$, we can write equations (*) as

$$\sigma\sigma^{-1} = \varepsilon \quad \text{and} \quad \sigma^{-1}\sigma = \varepsilon.$$

This and other properties of composition discussed earlier are recorded in the following theorem for reference.

Theorem 2. Let σ, τ , and μ denote permutations in S_n .

- (1) $\sigma\tau$ is in S_n .
- (2) $\sigma\varepsilon = \sigma = \varepsilon\sigma$.
- (3) $\sigma(\tau\mu) = (\sigma\tau)\mu$.
- (4) $\sigma\sigma^{-1} = \varepsilon = \sigma^{-1}\sigma$.

By virtue of this, S_n is said to be a *group under composition* that explains the name “symmetric group.” Groups in general are discussed in Chapter 2.

Example 4. Given

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix}$$

and

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 5 & 4 \end{pmatrix},$$

find χ in S_5 such that $\chi\sigma = \tau$.

Solution. Suppose that $\chi \in S_n$ exists such that $\tau = \chi\sigma$. Multiply on the right by σ^{-1} to get $\tau\sigma^{-1} = \chi\sigma\sigma^{-1} = \chi\varepsilon = \chi$. Thus,

$$\chi = \tau\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 5 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 5 & 4 & 3 & 2 \end{pmatrix}.$$

The reader should verify that χ actually works, that is, $\chi\sigma = \tau$. \square

Let $\sigma \in S_n$ so that $\sigma: X_n \rightarrow X_n$ is a bijection. We say that an element $k \in X_n$ is **fixed** by σ if $\sigma k = k$. If $\sigma k \neq k$, we say that k is **moved** by σ , and we write $M_\sigma = \{k \in X_n \mid k \text{ is moved by } \sigma\}$. Two permutations σ and τ are called **disjoint** if no element of X_n is moved by both; that is, if $M_\sigma \cap M_\tau = \emptyset$.

Clearly, the identity permutation ε in S_n is the only permutation that fixes every element of X_n . By contrast,

$$\begin{pmatrix} 1 & 2 & 3 & \cdots & n-1 & n \\ 2 & 3 & 4 & \cdots & n & 1 \end{pmatrix}$$

moves every element of X_n , whereas

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

moves 1, 3, and 5 and fixes 2 and 4. The following result is needed in the proof of Theorem 3.

Lemma 1¹⁴. *If $k \in M_\sigma$ then $\sigma k \in M_\sigma$.*

Proof. Otherwise, σk is fixed by σ ; that is, $\sigma(\sigma k) = \sigma k$. But then the fact that σ is one-to-one gives $\sigma k = k$, which is contrary to the hypothesis. \blacksquare

Theorem 3. *If σ and τ in S_n are disjoint, then $\sigma\tau = \tau\sigma$.*

Proof. For $k \in X_n$, we must show that $(\tau\sigma)k = (\sigma\tau)k$. Since $M_\sigma \cap M_\tau = \emptyset$ by hypothesis, there are three cases (see the diagram).

¹⁴The word “lemma” means a subsidiary proposition used in the proof of another proposition.

- *Case 1:* $k \in M_\sigma$. Then $\sigma k \in M_\sigma$ too (by Lemma 1), so neither lies in M_τ . Hence, both are fixed by τ , so $\tau k = k$ and $\tau(\sigma k) = \sigma k$. Hence,

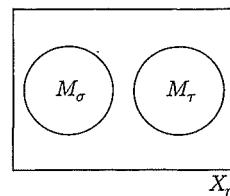
$$(\tau\sigma)k = \tau(\sigma k) = \sigma k = \sigma(\tau k) = (\sigma\tau)k.$$

- *Case 2:* $k \in M_\tau$. This case is analogous to Case 1, and is left to the reader.

- *Case 3:* $k \notin M_\sigma$ and $k \notin M_\tau$. Then $\sigma k = k$ and $\tau k = k$, so

$$(\tau\sigma)k = \tau(\sigma k) = \tau k = k = \sigma k = \sigma(\tau k) = (\sigma\tau)k.$$

This completes the proof. ■



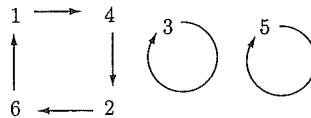
Note that the converse to Theorem 3 is not true. For example, $\sigma\sigma^{-1} = \sigma^{-1}\sigma$ for any σ in S_n , but σ and σ^{-1} are certainly not disjoint. Theorem 3 is important because it leads to a proof of the fact (Theorem 5 below) that every permutation in S_n can be written as a product of pairwise disjoint (and commuting) factors. We now turn our attention to this topic.

Cycles

Consider the permutation

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

in S_6 . The action of σ is described graphically as



Thus, the elements σ moves are moved in a cycle, and σ is called a *cycle* for this reason. We write σ as $\sigma = (1 \ 4 \ 2 \ 6)$. This notation lists only elements moved by σ , and each is moved to its neighbor to the right, except the last element, which “cycles around” to the first. We generalize this type of permutation as follows.

Let k_1, k_2, \dots, k_r be distinct elements of X_n .

Then, as shown in the diagram, the **cycle**

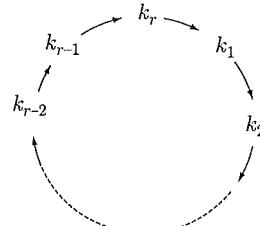
$$\sigma = (k_1 \ k_2 \ \dots \ k_r)$$

is the permutation in S_n defined by

$$\sigma k_i = k_{i+1}, \text{ if } 1 \leq i \leq r-1.$$

$$\sigma k_r = k_1$$

$$\sigma k = k, \quad \text{if } k \notin \{k_1, k_2, \dots, k_r\}$$



We say that σ has **length r** and refer to σ as an **r -cycle**. Note that the only cycle of length 1 is ε , that is $(k) = \varepsilon$ for each $k \in X_n$.

Example 5. Write

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 7 & 1 & 6 & 5 & 2 & 3 \end{pmatrix}$$

in cycle notation.

Solution. $\tau = (1 \ 4 \ 6 \ 2 \ 7 \ 3)$. Note that τ fixes 5. \square

Example 6. $S_3 = \{\varepsilon, (1 \ 2 \ 3), (1 \ 3 \ 2), (1 \ 2), (1 \ 3), (2 \ 3)\}$ from Example 1. Hence, S_3 consists of cycles; however, the same is not true of S_n in general, as we show later.

Example 7. The only cycle of length 1 is the identity permutation ε .

To reverse the action of a cycle, we simply go around the cycle in the opposite direction. Thus we obtain

Theorem 4. If σ is an r -cycle, then σ^{-1} is also an r -cycle. More precisely, if $\sigma = (k_1 \ k_2 \ \dots \ k_{r-1} \ k_r)$, then $\sigma^{-1} = (k_r \ k_{r-1} \ \dots \ k_2 \ k_1)$.

Cycle notation is much simpler than two-row matrix notation. However, we must briefly discuss two ambiguous aspects of cycle notation. First, the same permutation can be written in several ways in cycle notation. For example, $\sigma = (1 \ 4 \ 2 \ 3)$ in S_4 can be written as $\sigma = (4 \ 2 \ 3 \ 1) = (2 \ 3 \ 1 \ 4) = (3 \ 1 \ 4 \ 2)$. This is harmless once we are aware of it.

The second ambiguity can be illustrated as follows: Given $\sigma = (1 \ 2 \ 4)$, is it in S_4 (fixing 3) or in S_5 (fixing 3 and 5)? We introduce the following convention so that it does not matter.

Convention. Every permutation in S_n is regarded as a permutation in S_{n+1} that fixes $n+1$. Thus,

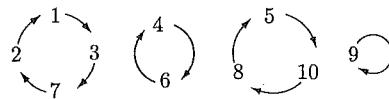
$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \dots$$

We shall adhere to this convention throughout this book.

Of course, not every permutation is a cycle. For example, consider

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 3 & 1 & 7 & 6 & 10 & 4 & 2 & 5 & 9 & 8 \end{pmatrix}$$

in S_{10} . If we represent the action of σ geometrically, we obtain



The four cycles are $(1 \ 3 \ 7 \ 2)$, $(4 \ 6)$, $(5 \ 10 \ 8)$, and $(9) = \varepsilon$. These are pairwise disjoint, so each commutes with the others by Theorem 3. Even more remarkable is the fact that σ is the product of these cycles (where we omit $(9) = \varepsilon$):

$$\sigma = (1 \ 3 \ 7 \ 2)(4 \ 6)(5 \ 10 \ 8).$$

The reader should check this assertion. In fact, every permutation can be expressed as a product of disjoint cycles in this way. Here is another example.

Example 8. Factor

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 \\ 5 & 12 & 2 & 1 & 9 & 11 & 4 & 3 & 7 & 10 & 13 & 8 & 6 \end{pmatrix}$$

as a product of (pairwise) disjoint cycles.

Solution. Starting with 1, follow the action of σ : $1 \rightarrow 5 \rightarrow 9 \rightarrow 7 \rightarrow 4 \rightarrow 1$. Thus, it has cycled, and the first cycle is $(1 \ 5 \ 9 \ 7 \ 4)$. Now start with any member of X_{13} not already considered, say $2 \rightarrow 12 \rightarrow 8 \rightarrow 3 \rightarrow 2$; so the next cycle is $(2 \ 12 \ 8 \ 3)$. However, 6 has still not been used. It provides the cycle $(6 \ 11 \ 13)$. The remaining member of X_{13} is 10 that is fixed by σ , so the corresponding cycle is $(10) = \varepsilon$. Hence,

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

is the desired factorization (where we drop the 1-cycles as before). Of course, the action of σ can be sketched as shown previously. \square

The method of Example 8 will express every permutation as a product of disjoint cycles because each cycle agrees with σ on the elements it moves, and these elements are fixed by the other cycles. In addition, the factorization is unique up to the order of the disjoint cycles, and we give a formal inductive proof of the following theorem at the end of this section.

Theorem 5. Cycle Decomposition Theorem. If $\sigma \neq \varepsilon$ is a permutation in S_n , then σ is a product of (one or more) disjoint cycles of length at least 2. This factorization is unique up to the order of the factors.

Example 9. List all the elements of S_4 , each factored into disjoint cycles.

Solution. The $4! = 24$ elements are as follows:

$$\begin{array}{lllll} \varepsilon & (1 \ 2) & (1 \ 2 \ 3) & (1 \ 2)(3 \ 4) & (1 \ 2 \ 3 \ 4) \\ & (1 \ 3) & (1 \ 2 \ 4) & (1 \ 3)(2 \ 4) & (1 \ 2 \ 4 \ 3) \\ & (1 \ 4) & (1 \ 3 \ 4) & (1 \ 4)(2 \ 3) & (1 \ 3 \ 2 \ 4) \\ & (2 \ 3) & (2 \ 3 \ 4) & & (1 \ 3 \ 4 \ 2) \\ & (2 \ 4) & (1 \ 3 \ 2) & & (1 \ 4 \ 2 \ 3) \\ & (3 \ 4) & (1 \ 4 \ 2) & & (1 \ 4 \ 3 \ 2) \\ & & (1 \ 4 \ 3) & & \\ & & (2 \ 4 \ 3) & & \end{array}$$

\square

The permutations in Example 9 are classified according to the following notion: Two permutations in S_n have the same **cycle structure** if, when they are factored into disjoint cycles, they have the same number of cycles of each length. We refer to this notation again later.

The Alternating Group

A cycle of length 2 is called a **transposition**. Thus, each transposition δ has the form $\delta = (m \ n)$ where $m \neq n$. Hence,

$$\delta^2 = \varepsilon \quad \text{and} \quad \delta^{-1} = \delta, \quad \text{for every transposition } \delta.$$

Note, however, that $\sigma = (1\ 2)(3\ 4)$ also satisfies $\sigma^2 = \varepsilon$ and $\sigma^{-1} = \sigma$, so these properties do not characterize the transpositions.

One reason for studying transpositions is that every permutation is a product of transpositions. For example, the cycle $(1\ 2\ 3\ 4\ 5\ 6)$ factors as follows:

$$(1\ 2\ 3\ 4\ 5\ 6) = (1\ 2)(2\ 3)(3\ 4)(4\ 5)(5\ 6)$$

as is easily verified. This pattern works in general.

Theorem 6. *Every cycle of length $r > 1$ is a product of $r - 1$ transpositions:*

$$(k_1\ k_2\ \cdots\ k_r) = (k_1\ k_2)(k_2\ k_3)\cdots(k_{r-2}\ k_{r-1})(k_{r-1}\ k_r).$$

Hence, every permutation is a product of transpositions.

Proof. The verification of the cycle factorization is left to the reader. The rest follows because every permutation is a product of cycles by Theorem 5. ■

In contrast to the factorization into cycles, factorizations into transpositions are *not* unique. For example,

$$(2\ 3)(1\ 2)(2\ 5)(1\ 3)(2\ 4) = (1\ 2\ 4\ 5) = (1\ 5)(1\ 4)(1\ 2).$$

Indeed, any factorization into m transpositions gives rise to a factorization into $m + 2$ transpositions simply by inserting $\varepsilon = (1\ 2)(1\ 2)$ somewhere. This gives a glimpse (admittedly not convincing!) into why the next theorem is true. It asserts that if a permutation can be factored in one way as a product of an even (or odd) number of transpositions, then *any* factorization into transpositions must involve an even (respectively odd) number of factors.

Two integers m and n are said to have the **same parity** if they are both even or both odd; equivalently, if $m \equiv n \pmod{2}$.

Theorem 7. Parity Theorem. *If a permutation σ has two factorizations*

$$\sigma = \gamma_n \cdots \gamma_2 \gamma_1 = \mu_m \cdots \mu_2 \mu_1,$$

where each γ_i and μ_j is a transposition, then m and n have the same parity.

The proof of this astonishing fact is given at the end of this section.

A permutation σ is called **even** or **odd** accordingly as it can be written in some way as the product of an even or odd number of transpositions. The parity theorem ensures that this is unambiguous, that is no permutation is both even and odd.

The parity of a cycle γ is easy to determine: Theorem 6 shows that γ is even if its length is odd, and odd if its length is even. When combined with Theorem 5, this result provides a way to easily compute the parity of any permutation.

Example 10. Determine the parity of $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 4 & 6 & 1 & 7 & 8 & 2 & 9 & 3 \end{pmatrix}$.

Solution. The factorization of σ into disjoint cycles is $\sigma = (1\ 5\ 7\ 2\ 4)(3\ 6\ 8\ 9)$. Then, $(1\ 5\ 7\ 2\ 4)$ is even and $(3\ 6\ 8\ 9)$ is odd by Theorem 6, so σ is odd (because the sum of an even and an odd integer is odd). □

The set of all even permutations in S_n is denoted A_n . It is called the **alternating group of degree n** and plays an important role in the theory of groups (in Chapter 2). Theorem 8 collects several facts about A_n that will be needed later.

Theorem 8. If $n \geq 2$, the set A_n has the following properties:

- (1) ε is in A_n and, if σ and τ are in A_n , then both σ^{-1} and $\sigma\tau$ are in A_n .
- (2) $|A_n| = \frac{1}{2}n!$

Proof. (1) $\varepsilon = (1 \ 2)(1 \ 2)$, so it is even. If σ and τ are even, write $\sigma = \gamma_1\gamma_2 \cdots \gamma_n$ and $\tau = \delta_1\delta_2 \cdots \delta_m$, where n and m are even and γ_i and δ_j are transpositions. Then $\sigma\tau = \gamma_1\gamma_2 \cdots \gamma_n\delta_1\delta_2 \cdots \delta_m$ is a product of $n+m$ transpositions, and so is even. Finally, write $\mu = \gamma_n \cdots \gamma_2\gamma_1$. The fact that $\gamma_i^2 = \varepsilon$ for each i implies that $\sigma\mu = \varepsilon$ (verify). Hence, $\sigma^{-1} = \sigma^{-1}\varepsilon = \sigma^{-1}\sigma\mu = \varepsilon\mu = \mu$. But μ is even because n is even, so σ^{-1} is even.

(2) Let O_n denote the set of odd permutations in S_n . Then $S_n = A_n \cup O_n$ and the parity theorem guarantees that $A_n \cap O_n = \emptyset$. Since $|S_n| = n!$, it suffices to show that $|A_n| = |O_n|$. We do so by exhibiting a bijection $f: A_n \rightarrow O_n$. Let $\gamma = (1 \ 2)$ and define f by $f(\sigma) = \gamma\sigma$ for all $\sigma \in A_n$. (Note that $\gamma\sigma$ is odd if σ is even.) The fact that $\gamma^2 = \varepsilon$ implies that f is a bijection. In fact, $\gamma\sigma = \gamma\sigma_1$ gives $\sigma = \gamma^2\sigma = \gamma^2\sigma_1 = \sigma_1$ (so f is one-to-one); if $\tau \in O_n$, then $\sigma = \gamma\tau \in A_n$ and $f(\sigma) = \gamma\sigma = \gamma^2\tau = \tau$ (so f is onto). Thus, $|A_n| = |O_n|$. ■

A set of permutations is called a *group* if it contains the identity permutation, the product of any two of its members, and the inverse of any member. Hence, S_n is a group, and the first part of Theorem 8 shows that A_n is a group. The general idea of a group is defined and discussed at length in Chapter 2.

Proof of the Cycle Decomposition Theorem

If $\sigma \neq \varepsilon$ is a permutation in S_n , we show it is a product of disjoint cycles by induction on $n \geq 2$. This is clear if $n = 2$. If $n > 2$, assume that the result is true for S_{n-1} and let $\sigma \in S_n$. If $\sigma n = n$, then $\sigma \in S_{n-1}$ and we are done. So assume $\sigma n \neq n$ and write $m = \sigma^{-1}n$. Then $\sigma m = \sigma(\sigma^{-1}n) = \varepsilon n = n$, and $m \neq n$ (because $\sigma n \neq n$). We write $\gamma = (m \ n)$ and consider $\tau = \sigma\gamma$. Because $\gamma^2 = \varepsilon$, we have $\tau\gamma = \sigma\gamma^2 = \sigma\varepsilon = \sigma$. Moreover, $\tau n = \sigma\gamma n = \sigma m = n$, so $\tau \in S_{n-1}$ and τ is a product of disjoint cycles by induction. There are two cases:

- *Case 1:* $\tau m = m$. In this case, γ and τ are disjoint (as $\tau n = n$) and we are done because $\sigma = \gamma\tau$.
- *Case 2:* $\tau m \neq m$. Then m is moved by (exactly one) cycle factor of τ . Hence we can write

$$\tau = \mu(m \ k_1 \ k_2 \ \cdots \ k_r),$$

where μ is a product of disjoint cycles fixing m, k_1, k_2, \dots, k_r (and also fixing n because $\tau n = n$). Finally, it is easy to verify that

$$\sigma = \tau\gamma = \mu(m \ k_1 \ k_2 \ \cdots \ k_r)(m \ n) = \mu(m \ n \ k_1 \ \cdots \ k_r),$$

which gives σ as a product of disjoint cycles.

Turning to the uniqueness, suppose that $\sigma = \gamma_a \dots \gamma_2 \gamma_1 = \delta_b \dots \delta_2 \delta_1$ are two factorizations into disjoint cycles. We proceed by induction on $\max(a, b)$. If this is 1, then $\sigma = \gamma_1 = \delta_1$. Otherwise, let σ move m . Then m occurs in exactly one γ_i and exactly one δ_j . By reordering the factors if necessary, assume that m occurs in γ_1 and in δ_1 . Hence, we can write

$$\gamma_1 = (k_1 \ k_2 \ \dots \ k_r) \quad \text{and} \quad \delta_1 = (l_1 \ l_2 \ \dots \ l_s),$$

where $k_1 = m = l_1$. We may assume that $r \leq s$. Then, because $k_1 = l_1$,

$$k_2 = \sigma k_1 = \sigma l_1 = l_2$$

$$k_3 = \sigma k_2 = \sigma l_2 = l_3$$

$$\vdots \qquad \vdots$$

$$k_r = \sigma k_{r-1} = \sigma l_{r-1} = l_r$$

If $r < s$, the next step gives

$$l_1 = k_1 = \sigma k_r = \sigma l_r = l_{r+1},$$

a contradiction. Thus, $r = s$ and $\gamma_1 = \delta_1$. If we write $\lambda = \gamma_1 = \delta_1$, we obtain $\sigma = \gamma_a \dots \gamma_2 \lambda = \delta_b \dots \delta_2 \lambda$. It follows that $\sigma \lambda^{-1} = \gamma_a \dots \gamma_2 = \delta_b \dots \delta_2$ is a product of $a - 1$ (and $b - 1$) disjoint cycles. By induction, $a = b$ and (after possible reordering) $\gamma_i = \delta_i$ for $i = 2, 3, \dots, a$, which completes the induction.

Proof of the Parity Theorem

The proof depends on two preliminary results about transpositions.

Lemma 2. *Let $\gamma_1 \neq \gamma_2$ be transpositions. If γ_1 moves k , transpositions δ_1 and λ_2 exist such that*

$$\gamma_2 \gamma_1 = \lambda_2 \delta_1, \quad \text{where } \delta_1 \text{ fixes } k \text{ and } \lambda_2 \text{ moves } k.$$

Proof. Let $\gamma_1 = (k \ a)$. Because $\gamma_1 \neq \gamma_2$, the transposition γ_2 has one of the forms $(k \ b)$, $(a \ b)$, or $(b \ c)$ where k, a, b , and c denote distinct integers. In these cases,

$$\gamma_2 \gamma_1 = (k \ b)(k \ a) = (k \ a)(a \ b)$$

$$\gamma_2 \gamma_1 = (a \ b)(k \ a) = (k \ b)(a \ b)$$

$$\gamma_2 \gamma_1 = (b \ c)(k \ a) = (k \ a)(b \ c)$$

Hence the conclusion of Lemma 2 holds in every case. ■

Lemma 3. *If the identity permutation ε can be written as a product of $n \geq 3$ transpositions, then it can be written as a product of $n - 2$ transpositions.*

Proof. Let $\varepsilon = \gamma_n \dots \gamma_4 \gamma_3 \gamma_2 \gamma_1$, where $n \geq 3$ and γ_i are transpositions. Suppose that γ_1 moves k . If $\gamma_1 = \gamma_2$, then $\gamma_2 \gamma_1 = \varepsilon$, so $\varepsilon = \gamma_n \dots \gamma_4 \gamma_3$ and we are done. Otherwise, Lemma 2 gives $\gamma_1 \gamma_2 = \lambda_2 \delta_1$, where δ_1 fixes k and λ_2 moves k . Thus,

$$\varepsilon = \gamma_n \dots \gamma_4 \gamma_3 \lambda_2 \delta_1.$$

Again, we are done if $\lambda_2 = \gamma_3$, so we let $\gamma_3 \lambda_2 = \lambda_3 \delta_2$, where δ_2 fixes k and λ_3 moves k . Hence,

$$\varepsilon = \gamma_n \dots \gamma_5 \gamma_4 \lambda_3 \delta_2 \delta_1.$$

Continue in this way. Either we are done at some stage or we finally arrive at a factorization

$$\varepsilon = \lambda_n \delta_{n-1} \cdots \delta_2 \delta_1,$$

where each δ_i fixes k and λ_n moves k . But this cannot happen because, if it did,

$$k = \varepsilon k = \lambda_n \delta_{n-1} \cdots \delta_2 \delta_1 k = \lambda_n k \neq k,$$

a contradiction. This proves Lemma 3. \blacksquare

Proof of the parity theorem. Suppose a permutation σ has two factorizations into transpositions:

$$\sigma = \gamma_n \cdots \gamma_2 \gamma_1 = \mu_m \cdots \mu_2 \mu_1.$$

We must show that n and m are both even or both odd. The fact that $\mu_j^{-1} = \mu_j$ for all j gives $\varepsilon = \mu_1 \mu_2 \cdots \mu_m \gamma_n \cdots \gamma_2 \gamma_1$. Hence, it suffices to show that ε cannot be written as the product of an odd number of transpositions. But if ε is a product of p transpositions, where $p \geq 3$ is odd, then repeating Lemma 3 gives factorizations into $p-2, p-4, \dots$, transpositions. Ultimately we get a factorization of ε as one transposition, which is impossible. \square

Exercises 1.4

1. Let

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 3 & 5 \end{pmatrix}, \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 5 & 4 \end{pmatrix}, \quad \mu = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix}$$

be permutations. Compute:

(a) $\tau\sigma$	(b) $\sigma\tau$	(c) τ^{-1}
(d) μ^{-1}	(e) $\mu\tau\sigma^{-1}$	(f) $\mu^{-1}\sigma\tau$

2. (a) Verify that any two of σ , τ , and μ commute:

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}, \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}, \quad \mu = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}.$$

(b) Do (a) by first verifying that $\sigma = \tau^2$ and $\mu = \tau^3$.

3. Let

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}$$

and

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix}.$$

In each case solve for χ in S_4 .

(a) $\sigma\chi = \tau$	(b) $\chi\tau = \sigma$	(c) $\sigma^{-1}\chi = \tau$
(d) $\chi\tau\sigma = \varepsilon$	(e) $\tau\chi\sigma = \varepsilon$	(f) $\tau\chi\sigma^{-1} = \sigma$

4. Suppose that

$$\tau\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 3 & 1 & 4 & 2 \end{pmatrix}$$

and

$$\sigma\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 5 & 1 \end{pmatrix}$$

in S_5 . If $\sigma 1 = 2$, find σ and τ .

5. Show that

$$\tau\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$$

and

$$\sigma\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

is impossible for σ and τ in S_4 .

6. If σ and τ fix k , show that $\sigma\tau$ and σ^{-1} both fix k .
7. (a) How many permutations in S_5 fix 1?
 (b) How many fix both 1 and 2?
8. (a) If $\sigma\tau = \varepsilon$ in S_n , show that $\sigma = \tau^{-1}$.
 (b) If $\sigma^2 = \sigma$ in S_n , show that $\sigma = \varepsilon$.
9. In S_n , show that $\sigma = \tau$ if and only if $\sigma\tau^{-1} = \varepsilon$.
10. If σ and τ are disjoint in S_n and $\sigma\tau = \varepsilon$, what can you say about σ and τ ? Support your answer.
11. Write the following in two-row matrix notation.
 (a) $(1 \ 8 \ 7 \ 4)(3 \ 6 \ 7 \ 5 \ 9)$ (b) $(1 \ 3 \ 5 \ 7)(4 \ 1 \ 9)$
12. Let $\sigma = (1 \ 2 \ 3)$ and $\tau = (1 \ 2)$ in S_3 .
 (a) Show that $S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$ and that $\sigma^3 = \varepsilon = \tau^2$ and $\sigma\tau = \tau\sigma^2$.
 (b) Use (a) to fill in the multiplication table for S_3 .
13. Factor each of the following permutations into disjoint cycles, find its parity, and factor the inverse into disjoint cycles.
 (a) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 4 & 7 & 9 & 8 & 2 & 1 & 6 & 3 & 5 \end{pmatrix}$
 (b) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 8 & 9 & 5 & 2 & 1 & 6 & 4 & 7 \end{pmatrix}$
 (c) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 8 & 6 & 9 & 4 & 7 & 3 & 1 & 5 \end{pmatrix}$
 (d) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 4 & 8 & 9 & 3 & 1 & 7 & 5 & 2 \end{pmatrix}$
 (e) $(1 \ 3)(2 \ 5 \ 7)(3 \ 8 \ 5)$
 (f) $(1 \ 2 \ 3 \ 4 \ 5)(6 \ 7)(1 \ 3 \ 5 \ 7)(1 \ 6 \ 3)$
14. If $\sigma\tau = \sigma\mu$ or $\tau\sigma = \mu\sigma$ in S_n , show that $\tau = \mu$. Does $\sigma\tau = \mu\sigma$ imply that $\tau = \mu$? Support your answer.
15. In each of (a) S_5 , and (b) S_6 , list one permutation of each possible cycle structure (see Example 9).
16. If $\sigma = (1 \ 2 \ 3 \ \cdots \ n)$, show that $\sigma^n = \varepsilon$ and that n is the smallest positive integer with this property.
17. (a) If $\sigma = (1 \ 2 \ 3 \ 4)(5 \ 6 \ 7)$, factor σ^{-1} into disjoint cycles.
 (b) If $\sigma = \gamma_1\gamma_2 \cdots \gamma_n$, where the γ_i are disjoint cycles, how is the factorization of σ^{-1} into disjoint cycles related to the γ_i ? Support your answer.

18. Find the parity of

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 5 & 11 & 6 & 1 & 15 & 13 & 2 & 9 & 4 & 10 & 14 & 3 & 12 & 7 & 8 \end{pmatrix}.$$

19. Find the parity of each permutation in Exercise 13.
20. Show that $(1 \ 2)$ is not a product of 3-cycles.
21. (a) If $\gamma_1, \gamma_2, \dots, \gamma_m$ are transpositions, show that
- $$(\gamma_1 \ \gamma_2 \ \cdots \ \gamma_m)^{-1} = \gamma_m \gamma_{m-1} \cdots \gamma_2 \gamma_1.$$
- (b) Show that σ and σ^{-1} have the same parity for all σ in S_n .
- (c) Show that σ and $\tau\sigma\tau^{-1}$ have the same parity for all σ and τ in S_n .
22. Show that $A_{n+1} \cap S_n = A_n$ for all $n \geq 3$ (regard $S_n \subseteq S_{n+1}$ in the usual way).
23. Let $\sigma \in S_n$, $\sigma \neq \epsilon$. If $n \geq 3$, show that $\gamma \in S_n$ exists such that $\sigma\gamma \neq \gamma\sigma$. [Hint: If $\sigma k = l$ with $k \neq l$, choose $m \notin \{k, l\}$ and take $\gamma = (k \ m)$.]
24. If $\sigma \in S_n$, show that $\sigma^2 = \epsilon$ if and only if σ is a product of disjoint transpositions.
25. If $n \geq 3$, show that every even permutation in S_n is a product of 3-cycles.
26. Let γ be any cycle of length r . If $\sigma \in S_n$, show that $\sigma\gamma\sigma^{-1}$ is also a cycle of length r . More precisely, if $\gamma = (k_1 \ k_2 \ \cdots \ k_r)$ show that $\sigma\gamma\sigma^{-1} = (\sigma k_1 \ \sigma k_2 \ \cdots \ \sigma k_r)$.
27. (a) Show that $(k_1 \ k_2 \ \cdots \ k_r) = (k_1 \ k_r)(k_1 \ k_{r-1}) \cdots (k_1 \ k_2)$.
- (b) Show that each $\sigma \in S_n$ is a product of the transpositions $(1 \ 2), (1 \ 3), \dots, (1 \ n)$. [Hint: Each transposition is such a product by (a) and Exercise 26.]
- (c) Repeat (b) for the transpositions $(1 \ 2), (2 \ 3), \dots, (n-1 \ n)$. [Hint: Use (a) and Exercise 26.]
- (d) If $\sigma = (1 \ 2 \ 3 \ \cdots \ n)$, show that each element of S_n is a product of the permutations $(1 \ 2), \sigma$, and σ^{-1} . [Hint: Use (b) and Exercise 26.]
28. Let $\sigma = (1 \ 2 \ 3 \ \cdots \ n)$ be a cycle of length $n \geq 2$.
- (a) If $n = 2k$, find the factorization of σ^2 into disjoint cycles.
- (b) If $n = mq$ with $m \geq 3$ and $q \geq 2$, show that σ^m is a product of m disjoint cycles, each of length q .
- (c) If $1 \leq m \leq n$, show that $\sigma^m k \equiv k + m \pmod{n}$.
- (d) If $n = p$ is a prime, show that σ^m is a cycle of length p for each $m = 1, 2, \dots, p-1$.
29. Define the sign of a permutation σ to be

$$\operatorname{sgn} \sigma = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd} \end{cases}.$$

Prove that $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn} \sigma \operatorname{sgn} \tau$ for all σ and τ in S_n .

30. Consider a puzzle made up of five numbered squares in a 2×3 frame. Assume that the squares slide vertically and horizontally so that rearrangements are possible. For example, arrangement (2) can be obtained from (1) (in four moves). Call an arrangement “nice” if the lower right position is vacant. Then, the “nice” arrangements correspond to permutations in S_5 . For example, arrangement (2) corresponds to $(2 \ 5 \ 3)$.

(1)	1	2	3	(2)	1	5	2
	4	5			4	3	

Show that every “nice” arrangement corresponds to an even permutation.¹⁵

¹⁵In fact, every even permutation arises in this way. (See Newman, J. R., *World of Mathematics*, New York: Simon & Schuster, 1956, p. 2431.)

1.5 AN APPLICATION TO CRYPTOGRAPHY

How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth.

—Sir Arthur Conan Doyle

The ability to transmit messages in a way that cannot be recognized by adversaries has intrigued people for centuries. In this brief section, we outline a method that uses Fermat's theorem to encode information in a way that is very difficult to break. The idea is based on the following consequence of that theorem.

Theorem 1. Let $n = pq$, where p and q are distinct primes, write $m = (p - 1)(q - 1)$, and let $e > 2$ be any integer such that $e \equiv 1 \pmod{m}$. Then

$$x^e \equiv x \pmod{n} \quad \text{for all } x \text{ such that } \gcd(x, n) = 1.$$

Proof. Because $e \equiv 1 \pmod{m}$, write $e - 1 = ym$, where y is an integer. Then $x^e = x \cdot (x^m)^y$, so it suffices to show that $x^m \equiv 1 \pmod{n}$ whenever $\gcd(x, n) = 1$. This condition certainly implies that p does not divide x . Hence, Fermat's theorem shows that $x^{p-1} \equiv 1 \pmod{p}$ and so $x^m = (x^{p-1})^{q-1} \equiv 1^{q-1} \equiv 1 \pmod{p}$. Similarly, $x^m \equiv 1 \pmod{q}$ and so, as p and q are relatively prime, Theorem 5 §1.2 shows that $x^m \equiv 1 \pmod{pq}$. This is what we wanted. ■

The coding process can be described as follows. Two distinct primes p and q are chosen, each very large in practice. Then the words available for transmission (and punctuation symbols) are paired with distinct integers $x \geq 2$. The integers x used may be assumed to be chosen relatively prime to p and q if these primes are large enough and, in practice, to be smaller than each of these primes. The idea is to use p and q to compute an integer r from x and then to transmit r rather than x . Clearly, r must be chosen in such a way that x (and hence the corresponding word) can be retrieved from r . The passage from x to r (called *encoding*) is carried out by the sender of a message, the integer r is transmitted, and the computation of x from r (*decoding*) is done by the receiver.

Here is how the process works. Given the distinct primes p and q , the cryptographer denotes

$$n = pq \quad \text{and} \quad m = (p - 1)(q - 1)$$

and then chooses any integer $k \geq 2$ such that $\gcd(k, m) = 1$. The sender is given only the numbers n and k . If the sender wants to transmit an integer x , he or she encodes it by reducing x^k modulo n , say,

$$x^k \equiv r \pmod{n}, \quad \text{where } 0 \leq r < n.$$

Then the sender transmits r to the receiver of the message who must use it to retrieve x . If the receiver knows the inverse k' of k in \mathbb{Z}_m , then $k'k \equiv 1 \pmod{m}$. Hence, Theorem 1 (with $e = k'k$) gives $x^{k'k} \equiv x \pmod{n}$ and

$$x \equiv x^{k'k} \equiv (x^k)^{k'} \equiv r^{k'}$$

modulo n . Knowing both r and k' , the receiver can compute x (and hence the corresponding word in the message).

Note that all the sender really has to know are n and k . A third party intercepting the message r cannot retrieve x without k' , and computing it requires p and q .

Even if the third party can extract the integers n and k from the sender, factoring $n = pq$ in practice is very time-consuming if the primes p and q are large, even with a computer. Hence, the code is extremely difficult to break. Example 1 illustrates how the process works, although the primes used are small.

Example 1. Let $p = 11$ and $q = 13$ so that $n = 143$ and $m = 120$. Then let $k = 7$, chosen so that $\gcd(k, m) = 1$. Encode the number $x = 9$ and then decode the result.

Solution. The sender reduces $x^k = 9^7$ modulo $n = 143$. Working modulo 143: $9^2 \equiv 81$, $9^3 \equiv 14$, $9^4 \equiv 126$, $9^7 \equiv 48$. Hence, $r = 48$ is transmitted. The receiver then finds k' , the inverse of $k = 7$ modulo $m = 120$. In fact, the Euclidean algorithm gives $1 = 120 - 17 \cdot 7$, so $k' \equiv -17 \equiv 103 \pmod{120}$ is the required inverse. Hence, x is retrieved (modulo n) by $x \equiv r^{k'} \equiv 48^{103} \pmod{143}$. One fairly efficient way to compute this is to note that $103 = 1100111$ in binary, so $103 = 1 + 2 + 2^2 + 2^5 + 2^6$. Then the receiver computes 48^t , where t is a power of 2 by successive squaring of 48 modulo 143:

$$48^2 \equiv 16, 48^3 \equiv 113, 48^4 \equiv 42, 48^5 \equiv 48, 48^6 \equiv 16, 48^7 \equiv 113.$$

Again working modulo 143 gives

$$x \equiv 48^{103} \equiv 48^{1+2+2^2+2^5+2^6} \equiv 48 \cdot 16 \cdot 113 \cdot 16 \cdot 113 \equiv 9,$$

which retrieves the original 9. □

This system is called the RSA system after its inventors.¹⁶ Other, more comprehensive coverage of cryptography is available,¹⁷ including overviews of the subject, methods, and bibliographies.

The RSA system works by finding two large primes p and q and computing the number $n = pq$. The code is difficult to break because it is difficult to find p and q given n . However, in 2002, Manindra Agrawal and two undergraduate students (Neeraj Kayal and Nitin Saxena) gave a simple algorithm that can decide whether a given integer n is prime or not. Moreover, the time taken is approximately a polynomial function of n . This is an important breakthrough in computer science, and certainly affects algorithms like the RSA system.

Cryptography, in general, refers to the transmission of messages where the primary aim is to disguise the message to make its interpretation by an unauthorized interceptor very difficult. Coding theory, in contrast, aims at fast and correct transmission of messages; we briefly discuss this topic in Sections 2.11 and 6.7.

¹⁶Rivest, R. L., Shamir, A., and Adleman, L., A method for obtaining digital signatures and public-key cryptosystems, *Communication of the ACM*, 21 (1978), 120–126.

¹⁷For example, see the section on Algebraic Cryptography in Lidl, R. and Pilz, G., *Applied Abstract Algebra*, New York: Springer-Verlag, 1983.

Chapter 2

Groups

Wherever groups disclose themselves, or could be introduced, simplicity crystallizes out of complete chaos.

—Eric Temple Bell

The origin of the modern theory of groups lies in the theory of equations. By the beginning of the nineteenth century, mathematicians had developed formulas for finding the roots of any cubic or quartic equation (analogous to the quadratic formula), and the best mathematicians of the day were trying to find such a formula for the quintic. It thus came as a great surprise when, in 1824, Niels Henrik Abel proved that no such formula exists. At about the same time, Evariste Galois showed that any equation of degree n has an associated group of permutations of the roots of the equation (that is, a set of permutations closed under compositions and inverses). He proved that the equation is solvable if and only if this group has a certain property (now called a *solvable group*). In particular, the fact that the group A_n of even permutations is not solvable for any $n \geq 5$ implies that no formula exists for solving equations of degree $n \geq 5$. This spectacular achievement led to modern Galois theory, but Galois' work went unrecognized until after his death at age 20.

Galois worked with groups of permutations. Then, in 1854, Arthur Cayley formulated the abstract group concept. While the study of permutation groups continues to occupy mathematicians, the abstract theory has the advantage that it isolates those properties of groups that do not depend on the underlying permutations and so can be applied more broadly. We pursue the abstract theory in this chapter (and in Chapters 7–9) with permutation groups as one of the most important examples.

2.1 BINARY OPERATIONS

Abstract algebra is primarily concerned with the study of operations analogous to the addition and multiplication of numbers. We define such operations in this section and examine some of their general properties. The addition process for numbers assigns to any ordered pair (a, b) of numbers a new number, their sum, denoted $a + b$. Similarly, multiplication assigns the product ab to the pair (a, b) .

In general, a **binary operation** $*$ on a set M is a mapping that assigns to each ordered pair (a, b) of elements of M an element $a * b$ of M . In this case M is said to be **closed** under the binary operation. Binary operations are usually denoted by other symbols (for example, $+$ for numbers) but, for the moment, we use the generic notation $a * b$.

A binary operation $*$ is called **commutative** if $a * b = b * a$ for all a, b in M , and $*$ is called **associative** if $a * (b * c) = (a * b) * c$ for all a, b, c in M . An element e in M is called a **unity** (or an **identity**)¹⁸ for $*$ if $a * e = a = e * a$ for all a in M . The unity for a binary operation is often denoted by different symbols (for example, 0 and 1 are the identities for addition and multiplication of numbers, respectively).

Theorem 1. If a binary operation has a unity, that unity is unique.

Proof. If e and f are both unities, then $f = e * f$ and $e * f = e$. So $e = f$. ■

A set M is called a **monoid** if a binary operation is defined on M that is associative and has an unity¹⁹. We say that $(M, *)$ is a monoid if the operation $*$ is to be emphasized. If the operation is commutative, we say that M is a **commutative monoid**.

Example 1. The sets \mathbb{Z} , \mathbb{Q} , \mathbb{R} , \mathbb{C} , and \mathbb{Z}_n are all commutative monoids under both addition and multiplication. The additive unity is 0 in all cases ($\bar{0}$ in \mathbb{Z}_n), and the multiplicative unity is 1 ($\bar{1}$ in \mathbb{Z}_n).

Example 2. The set $M_n(\mathbb{R})$ of all $n \times n$ real matrices is a monoid under both matrix addition and matrix multiplication, the unities being 0 and I , respectively. The monoid $(M_n(\mathbb{R}), +)$ is commutative. However, $(M_n(\mathbb{R}), \cdot)$ is *not* commutative if $n \geq 2$ (the proof of associativity is given in Appendix B).

Example 3. If U is a set, let $M = \{X \mid X \subseteq U\}$ denote the set of all subsets of U . Then (M, \cup) and (M, \cap) are both commutative monoids, the unities being \emptyset and U , respectively.

Example 4. S_n is a monoid with unity ε , and it is noncommutative if $n \geq 3$ (see Exercise 23 §1.4).

Example 5. If X is a nonempty set, let $M = \{\alpha \mid \alpha : X \rightarrow X \text{ is a mapping}\}$. Then M is a monoid using composition of mappings as the operation and the identity mapping 1_X as the unity (Theorem 3 §0.3). Moreover, M is noncommutative if X has at least two elements.

¹⁸The term “identity” is often used here but it has other meanings in algebra. So we use the term **unity**.

¹⁹A set with an associative binary operation, but possibly no unity, is called a *semigroup*.

Example 6. Let $*$ be the operation defined on \mathbb{N} by $n * m = n^m$. This operation is neither commutative ($2 * 3 = 8$ but $3 * 2 = 9$) nor associative ($(2 * 3) * 2 = 64$, but $2 * (3 * 2) = 512$), and there is no unity ($m = x * m$ for all m is impossible). Thus $(\mathbb{N}, *)$ is not a monoid. Note, however, that $m * 1 = m$ for all m .

A comment on notation is in order here. Binary operations are denoted by many different symbols in mathematics. For example, $+$ and \cdot are universally used for addition and multiplication of numbers, but these symbols are also standard for the addition and multiplication of matrices. Similarly, \cap and \cup are well-established notations in set theory. When a binary operation has such a standard symbol, we use it along with any standard notation for the corresponding unity (as in the foregoing examples). However, when discussing monoids *in general*, we have been using $*$ for the binary operation. But algebraists do not do this. They usually adopt one of the following two formats.

- **Multiplicative Notation.** Here $a * b$ is written as ab (or sometimes $a \cdot b$) and is called the *product* of a and b . The multiplicative unity is denoted 1 (or 1_M if the monoid M must be emphasized).
- **Additive Notation.** Here $a * b$ is written as $a + b$ and is called the *sum* of a and b . The additive unity is denoted 0 (or 0_M if the monoid M must be emphasized).

Multiplicative notation is the most popular format among algebraists. Hence we adopt the following convention.

Convention. In dealing with monoids *in general*, we use multiplicative notation, and denote the unity by 1 .

Hence ab can mean many different things, depending on the monoid under discussion, but the meaning is nearly always clear from the context. The small amount of confusion is more than balanced by the simplicity and conciseness of the notation.

For a finite monoid M , defining the operation by means of a table is sometimes convenient (as in Example 7 below). Given x and y in M , the product xy is the entry of the table in the row corresponding to x and the column corresponding to y . Hence, for the table in Example 7, $ab = b$ and $ca = e$. The elements of the monoid appear in the same order across the top of the table as down the left side. Such a table is called the **Cayley table** of the monoid, honoring Arthur Cayley who used it in 1854.

Example 7. If $M = \{e, a, b, c\}$, consider the binary operation shown in the table. The first row and column show that e is the

unity. That the operation is commutative is also

clear from the table because the entries are sym-

metric about the main diagonal (upper left to lower

right). However, this operation is not associative.

For example $a(bc) = ac = e$ while $(ab)c = bc = c$.

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

If a, b, c , and d are elements in a monoid M , there are various ways to form the product $abcd$ —for example $[(ab)c]d$ and $a[b(cd)]$. Verifying that these forms are equal is not difficult using associativity. In fact, we have

Theorem 2. General Associativity. Let a_1, a_2, \dots, a_n be elements of a monoid M . If the product $a_1 a_2 \cdots a_n$ is formed (in that order), the result is the same no matter which bracketing is used.

*Proof*²⁰. Let the *standard product* $\langle a_1, a_2, \dots, a_n \rangle$ be defined inductively by setting $\langle a_1 \rangle = a_1$ and $\langle a_1, a_2, \dots, a_n \rangle = a_1 \langle a_2, \dots, a_n \rangle$ for $n \geq 2$. Thus, $\langle a_1, a_2 \rangle = a_1 a_2$, $\langle a_1, a_2, a_3 \rangle = a_1(a_2 a_3)$, and so on. We use strong induction on $n \geq 1$ to prove the following statement: *If p is any product of a_1, a_2, \dots, a_n in that order, then $p = \langle a_1, a_2, \dots, a_n \rangle$.* This is clear if $n = 1$ or $n = 2$; if $n = 3$, the only non-standard product is $(a_1 a_2) a_3$, which equals $\langle a_1, a_2, a_3 \rangle = a_1(a_2 a_3)$ by associativity. In general, because p is formed using multiplication, it must factor as $p = qr$, where q is a product of a_1, a_2, \dots, a_k and r is a product of a_{k+1}, \dots, a_n for some k with $1 \leq k \leq n - 1$. Hence $r = \langle a_{k+1}, \dots, a_n \rangle$ by induction. If $k = 1$, then

$$p = a_1 \langle a_2, \dots, a_n \rangle = \langle a_1, a_2, \dots, a_n \rangle$$

as required. If $k > 1$, then $q = \langle a_1, \dots, a_k \rangle = a_1 \langle a_2, \dots, a_k \rangle$ by induction, and

$$\begin{aligned} p &= (a_1 \langle a_2, \dots, a_k \rangle) \langle a_{k+1}, \dots, a_n \rangle \\ &= a_1 (\langle a_2, \dots, a_k \rangle \langle a_{k+1}, \dots, a_n \rangle) \quad (\text{by associativity}) \\ &= a_1 \langle a_2, \dots, a_n \rangle \quad (\text{by induction}) \\ &= \langle a_1, a_2, \dots, a_n \rangle \end{aligned}$$

which completes the proof. ■

Theorem 2 enormously simplifies notation. It means that, in a monoid, we may (and do) write $a_1 a_2 \cdots a_n$ for the product of n elements with no ambiguity. If the operation were not associative, we would have to be careful about which bracketing we use. Of course, the *order* of the factors in a product does make a difference if the operation is not commutative.

Let a be an element of a monoid M . If $n \geq 0$ is an integer, inductively define the *n th power* a^n of a as follows:

$$a^0 = 1; \quad a^n = a \cdot a^{n-1}, \quad \text{for all } n \geq 1.$$

Thus, $a^1 = a$, $a^2 = a \cdot a$, $a^3 = a \cdot a \cdot a$, and so on. The following laws, familiar for numbers, hold for any monoid.

Theorem 3. Exponent Laws. Let a and b be elements of a monoid M .

- (1) $a^n a^m = a^{n+m}$ for all $n \geq 0$ and $m \geq 0$.
- (2) $(a^n)^m = a^{nm}$ for all $n \geq 0$ and $m \geq 0$.
- (3) If $ab = ba$, then $(ab)^n = a^n b^n$ for all $n \geq 0$.

Proof. (1) Fix $m \geq 0$ and prove (1) by induction on $n \geq 0$. If $n = 0$ then $a^0 a^m = 1 a^m = a^m = a^{0+m}$. If $n \geq 1$ then $a^n a^m = (aa^{n-1})a^m = a(a^{n-1}a^m)$. Since $a^{n-1}a^m = a^{n-1+m}$ by induction, this gives $a^n a^m = a(a^{n-1+m}) = a^{n+m}$.

²⁰This proof will not be used below and so may be omitted at a first reading. By contrast, the theorem will be used hundreds of times.

(2) Fix $n \geq 0$ and induct on m using (1). If $m = 0$, then $(a^n)^0 = 1 = a^{n \cdot 0}$. If $m \geq 1$, then $(a^n)^m = a^n \cdot (a^n)^{m-1} = a^n a^{n(m-1)} = a^{n+n(m-1)} = a^{nm}$.

(3) This assertion follows by induction on n after first showing $ba^n = a^n b$ for all $n \geq 0$ (Exercise 10). \blacksquare

It is interesting to note that, in the monoids of Example 3 (with \cap and \cup as the operations), $a^2 = a$ for all a . Hence $a^n = a$ for all $n \geq 1$.

Inverses

If s is a nonzero real number, the inverse $\frac{1}{s}$ is the solution to the equation $xs = 1$. In this form the idea extends to any monoid. If a is an element in a monoid M , an element b of M is called an **inverse** of a if $ab = 1 = ba$. An element with an inverse is called a **unit**. Note that the definition is symmetric in a and b , so that a is an inverse of b if and only if b is an inverse of a .

Theorem 4. *If M is a monoid and $a \in M$ has an inverse in M , then that inverse is unique.*

Proof. If both b and b' are inverses of a , then $ab = 1 = ba$ and $ab' = 1 = b'a$. Hence $b' = b'1 = b'(ab) = (b'a)b = 1b = b$. \blacksquare

Note the use of associativity in Theorem 4. In fact, its use is essential: In Example 7, both a and c are inverses of c .

If a is a unit in a multiplicative monoid, the inverse of a is denoted a^{-1} . If the monoid is additive the inverse of a is denoted $-a$ and is called the **negative** of a .

Example 8. Consider the additive monoids $(\mathbb{Z}, +)$, $(\mathbb{R}, +)$, $(\mathbb{C}, +)$, $(\mathbb{Z}_n, +)$, and $(M_n(\mathbb{R}), +)$. Then every element is a unit and, in all cases, the usual negative $-x$ of an element x is the (additive) inverse.

Example 9. In the multiplicative monoids (\mathbb{R}, \cdot) and (\mathbb{C}, \cdot) , every nonzero element is a unit. However, 0 has no inverse in either case.

Example 10. The units of (\mathbb{Z}, \cdot) are 1 and -1 .

Example 11. The units in (\mathbb{Z}_n, \cdot) are the residues \bar{a} , where a and n are relatively prime (Theorem 5 §1.3).

Example 12. If $M = \{\alpha \mid \alpha : X \rightarrow X \text{ is a mapping}\}$ under composition, the units in M are the bijections (onto and one-to-one mappings). (See Theorem 6 §0.3.)

Example 12 is important, and we refer to it again. If $X = \{1, 2, \dots, n\}$, the set of units is the symmetric group S_n of degree n (Section 1.4). If $X = \mathbb{N}$, we get a monoid containing maps σ and τ such that $\sigma\tau = 1$ but $\tau\sigma \neq 1$. (Example 8 §0.3.)

Example 13. The units in $(M_n(\mathbb{R}), \cdot)$ are the matrices A with $\det A \neq 0$, where $\det A$ denotes the determinant of A .²¹ This is discussed in Appendix B.

²¹See Nicholson, W.K., *Linear Algebra with Applications*, 7th ed., McGraw-Hill Ryerson, 2012 (Theorem 2 §3.2).

The next theorem collects several basic properties of units that will be used without comment throughout the book. This theorem will be familiar to students of linear algebra where it is proved for invertible matrices.

Theorem 5. Let $a, b, a_1, a_2, \dots, a_{n-1}, a_n$ denote elements in a monoid M .

- (1) 1 is a unit and $1^{-1} = 1$.
 - (2) If a is a unit so is a^{-1} , and $(a^{-1})^{-1} = a$.
 - (3) If a and b are units so is ab , and $(ab)^{-1} = b^{-1}a^{-1}$.
 - (4) If $a_1, a_2, \dots, a_{n-1}, a_n$ are units, so is $a_1a_2 \cdots a_{n-1}a_n$, and
- $$(a_1a_2 \cdots a_{n-1}a_n)^{-1} = a_n^{-1}a_{n-1}^{-1} \cdots a_2^{-1}a_1^{-1}.$$
- (5) If a is a unit so is a^n for any $n \geq 0$, and $(a^n)^{-1} = (a^{-1})^n$.

Proof. (1), (2), and (3) depend on the fact that, if $ab = 1 = ba$ for some b then a is a unit and $a^{-1} = b$. Thus (1) follows from $1 \cdot 1 = 1$; (2) follows from $a^{-1}a = 1 = aa^{-1}$, and (3) follows if we can show that

$$(ab)(b^{-1}a^{-1}) = 1 \quad \text{and} \quad 1 = (b^{-1}a^{-1})(ab).$$

But $(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = a1a^{-1} = aa^{-1} = 1$. The other equation can be similarly verified.

Finally (4) follows from (3) by induction on n (Exercise 16), and (5) is the special case of (4) where $a_1 = a_2 = \cdots = a_n = a$. ■

Note that every monoid has at least one unit: the unity. Moreover, if M is the set of all subsets of a set U , then (M, \cap) and (M, \cup) are monoids in which the unity is the *only* unit. At the other extreme are the monoids (called groups) in which *every* element is a unit. These are the principal objects of study in this chapter. With this in mind, we extend the definition of n th powers to include negative powers of a unit. Since $(a^{-1})^n = (a^n)^{-1}$ by Theorem 5 for any unit a , we define the negative powers a^{-n} , $n \geq 1$, by

$$a^{-n} = (a^{-1})^n = (a^n)^{-1}.$$

Then the laws of exponents extend as follows (the proof is left to the reader).

Theorem 6. Let a and b denote units in a monoid M .

- (1) $a^n a^m = a^{n+m}$ for all $n, m \in \mathbb{Z}$.
- (2) $(a^n)^m = a^{nm}$ for all $n, m \in \mathbb{Z}$.
- (3) If $ab = ba$, then $(ab)^n = a^n b^n$ for all $n \in \mathbb{Z}$.

Exercises 2.1

1. In each case a binary operation $*$ is given on a set M . Decide whether it is commutative or associative, whether a unity exists, and find the units (if there is a unity).
 - (a) $M = \mathbb{Z}$; $a * b = a - b$
 - (b) $M = \mathbb{Q}$; $a * b = \frac{1}{2}ab$
 - (c) $M = \mathbb{R}$; $a * b = a + b - ab$
 - (d) $M = \text{any set with } |M| \geq 2$; $a * b = b$

- (e) $M = P \times Q$, where P and Q are sets with $|P| \geq 2$ and $|Q| \geq 2$;
 $(p, q) * (p', q') = (p, q')$
- (f) $M = \mathbb{N}$; $m * n = \max(m, n)$ —the larger of m and n
- (g) $M = \mathbb{N}^+$; $a * b = \gcd(a, b)$
- (h) $M = \mathbb{R} \times \mathbb{R}$; $(x, y) * (x', y') = (xx', xy')$
- (i) $M = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$; $(x, y, z) * (x', y', z') = (xx', xy' + yz', zz')$
- (j) $M = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$; $(x, y, z) * (x', y', z') = (xy', yy', yz')$
2. (a) If x, y , or z is 1, show that $(xy)z = x(yz)$.
(b) Show that there are exactly two monoids with two elements.
(c) Let S be a set with an associative binary operation but with no unity. Choose an element $1 \notin S$, write $M = \{1\} \cup S$, and define an operation on M by using the operation of S and $1s = s = s1$ for all $s \in S$. Show that M is a monoid.
3. Consider the partial Cayley tables: (1)

	a	b
a		b
b		a

 and (2)

	a	b
a		a
b		b
- (a) Show that there is only one way to complete table (1) so that the resulting operation is associative, and that the result makes $\{a, b\}$ into a commutative monoid.
(b) Show that there are three associative completions of table (2), two making $\{a, b\}$ into a commutative monoid and one having no unity.
4. If M is any monoid, let \bar{M} denote the set of all nonempty subsets of M and define an operation on \bar{M} by $XY = \{xy \mid x \in X, y \in Y\}$. Show that \bar{M} is a monoid, commutative if M is, and find the units.
5. Given an alphabet A , call an n -tuple (a_1, a_2, \dots, a_n) with $a_i \in A$ a **word of length n** from A and write it (as in English) as $a_1 a_2 \dots a_n$. Multiply two words by

$$(a_1 a_2 \dots a_n) \cdot (b_1 b_2 \dots b_m) = a_1 a_2 \dots a_n b_1 b_2 \dots b_m,$$
- and call this product **juxtaposition**. Thus, the product of “no” and “on” is “noon”. We decree the existence of an **empty word** λ with no letters. Show that the set W of all words from A is a monoid, noncommutative if $|A| > 1$, and find the units.
6. Given a set X and a monoid M , let $F = \{\sigma \mid \sigma : X \rightarrow M \text{ is a mapping}\}$. Given σ and τ in F , define $\sigma \cdot \tau : X \rightarrow M$ by $(\sigma \cdot \tau)(x) = \sigma(x)\tau(x)$. Show that this definition makes F into a monoid, commutative if M is, and find all the units.
7. If M and N are monoids, show that the cartesian product $M \times N$ is a monoid (called the **direct product** of M and N) using the operation $(m, n)(m', n') = (mm', nn')$. When is $M \times N$ commutative? Describe the units.
8. An element e of a monoid M is called an **idempotent** if $e^2 = e$.
- (a) If $a \in M$ satisfies $a^m = a^{m+n}$, where $m \geq 0$ and $n \geq 1$, show that some power of a is an idempotent. [Hint: $a^{m+r} = a^{m+kn+r}$ for all $k \geq 1$ and $r \geq 0$.]
(b) If M is finite, show that some positive power of every element is an idempotent.
9. Assume that a is left cancelable in a monoid M ($ab = ac$ implies that $b = c$).
(a) If $a^5 = b^5$ and $a^{12} = b^{12}$ in M , show that $a = b$.
(b) If $a^m = b^m$ and $a^n = b^n$ where m and n are relatively prime, show that $a = b$.
10. If $ab = ba$ in a monoid M , prove that $(ab)^n = a^n b^n$ for all $n \geq 0$ (Theorem 3(3)).
11. An element e is called a **left (right) unity** for an operation if $ex = x$ ($xe = x$) for all x . If an operation has two left unities, show that it has no right unity.
12. (a) If u is a unit in a monoid M , show that $au = bu$ in M implies that $a = b$.
(b) If M is a finite monoid and $au = bu$ in M implies that $a = b$, show that u is a unit. [Hint: If $M = \{a_1, \dots, a_n\}$ show that $a_1 u, \dots, a_n u$ are distinct.]

13. If uv is a unit in a monoid M , and if $av = bv$ implies that $a = b$ in M , show that u and v are both units.
14. If $uv = 1$, we say that u is a **left inverse** of v and v is a **right inverse** of u . If u has both a left and a right inverse in a monoid, show that u is a unit.
15. If M is a monoid and $u \in M$, let $\sigma : M \rightarrow M$ be defined by $\sigma(a) = ua$ for all $a \in M$.
- Show that σ is a bijection if and only if u is a unit,
 - If u is a unit, describe the inverse mapping $\sigma^{-1} : M \rightarrow M$.
16. If $u_1, u_2, \dots, u_{n-1}, u_n$ are units in a monoid, show that $u_1 u_2 \cdots u_{n-1} u_n$ is also a unit and that $(u_1 u_2 \cdots u_{n-1} u_n)^{-1} = u_n^{-1} u_{n-1}^{-1} \cdots u_2^{-1} u_1^{-1}$ (Theorem 5(4)).
17. Let u and v be units in a monoid M .
- If $u^{-1} = v^{-1}$, show that $u = v$.
 - If $a \in M$ and $ua = au$, show that $u^{-1}a = au^{-1}$.
 - If $uv = vu$, show that $u^{-1}v^{-1} = v^{-1}u^{-1}$.
18. Prove that the following are equivalent for a monoid M .
- If ab is a unit then both a and b are units.
 - If $ab = 1$, then $ba = 1$.
19. If M is a finite monoid and $uv = 1$ in M , prove that $vu = 1$. [Hint: Exercise 12(b).]
20. Let M be a commutative monoid. Define a relation \sim on M by $a \sim b$ if $a = bu$ for some unit u .
- Show that \sim is an equivalence on M .
 - If \bar{a} denotes the equivalence class of a , let $\bar{M} = \{\bar{a} \mid a \in M\}$ denote the set of all equivalence classes. Show that $\bar{a}\bar{b} = \bar{a}\bar{b}$ is a well-defined operation on \bar{M} .
 - If \bar{M} is as in (b), show that \bar{M} is a commutative monoid in which the unity $\bar{1}$ is the only unit.
21. If M is a monoid, define $E(M) = \{\alpha : M \rightarrow M \mid \alpha(xy) = \alpha(x) \cdot y \text{ for all } x, y \in M\}$. If $a \in M$, define $\alpha_a : M \rightarrow M$ by $\alpha_a(x) = ax$ for all $x \in M$.
- Show that $E(M)$ is a monoid under composition of mappings.
 - Show that $\alpha_a \in E(M)$ for all $a \in M$.
 - If $\theta : M \rightarrow E(M)$ is defined by $\theta(a) = \alpha_a$ for all $a \in M$, show that θ is onto and one-to-one, $\theta(1) = 1_M$, and $\theta(ab) = \theta(a) \cdot \theta(b)$ for all $a, b \in M$.
22. Show that there are exactly six monoids M with three elements. If $M = \{1, a, b\}$, consider first the case $a^2 = 1$ (then only one multiplication table is possible). If $a^2 = b$, then $M = \{1, a, a^2\}$ is commutative and there are three monoids. Then two more emerge if $a^2 = a$. Note that, although associativity is used to force the multiplication table in every case, the associativity in the resulting table must be *checked* (Exercise 2(a) is useful).

2.2 GROUPS

A group is a monoid in which every element has an inverse. Because of its importance, we give the definition in full detail. A set G is called a **group** if it satisfies the following axioms.

- G1 G is closed under a binary operation.
- G2 The operation is associative.
- G3 There is a unity element in G .
- G4 Every element of G has an inverse in G .

The group G is called **abelian**²² if, in addition, it satisfies

G5 *The operation is commutative.*

If G is finite, the number $|G|$ of elements in G is called the **order** of G .

The terminology used for groups (and other algebraic systems, such as monoids) is somewhat careless. Strictly speaking, a group (G, \cdot) consists of *two* things: a set G and a binary operation. However, common practice is simply to refer to a group G and not mention the operation. This practice usually causes no difficulty, because the operation in the group in question is understood. We adopt this loose notation because it is much simpler, and also to acquaint the reader with what is in fact used in more advanced books. When clarity is needed, we use terms such as the group $(G, +)$ or the additive group G .

Examples 1–10 indicate the variety of ways that groups can occur, and we refer to many of them later. We leave verification of the axioms to the reader.

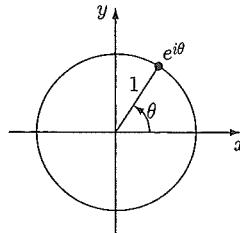
Example 1. $\{1\}$, $\{1, -1\}$, and $\{1, -1, i, -i\}$ are all abelian groups of (complex) numbers under multiplication. Here -1 is self-inverse, and i and $-i$ are inverses of each other.

Example 2. $\mathbb{Q} \setminus \{0\}$,²³ $\mathbb{R} \setminus \{0\}$, and $\mathbb{C} \setminus \{0\}$ are all abelian groups under multiplication. In each case the inverse of an element a is $a^{-1} = 1/a$.

Example 3. The set of complex numbers

$$\mathbb{C}^0 = \{z \in \mathbb{C} \mid |z| = 1\} = \{e^{i\theta} \mid \theta \in \mathbb{R}\}$$

is a group under complex multiplication. Here $e^{i\theta} = \cos \theta + i \sin \theta$, as in Appendix A, and we have $e^{i\theta} e^{i\varphi} = e^{i(\theta+\varphi)}$ and $(e^{i\theta})^{-1} = e^{-i\theta}$. The group \mathbb{C}^0 is called the **circle group** because it consists of the points on the unit circle.



Example 4. For $n \geq 1$, the group $U_n = \{z \in \mathbb{C} \mid z^n = 1\}$ is a group under complex multiplication, called the group of *n*th **roots of unity**. As in Appendix A, we have

$$U_n = \{e^{2k\pi i/n} \mid k = 0, 1, 2, \dots, n-1\}.$$

Clearly, $U_n \subseteq \mathbb{C}^0$ for all $n \geq 1$, and U_1 , U_2 , and U_4 are displayed in Example 1.

Example 5. The sets \mathbb{Z} , \mathbb{Q} , \mathbb{R} , and \mathbb{C} are all abelian groups under addition. In each case the unity is 0 and the inverse of x is $-x$.

Although we write most groups multiplicatively, many important groups are written additively (as in Example 5). Then the unity element is denoted 0 and is called **zero**, and the inverse of x is denoted $-x$ and is called the **negative** of x .

²²The name honors the Norwegian mathematician Niels Henrik Abel.

²³If X and Y are sets the **set difference** is defined by $X \setminus Y = \{x \in X \mid x \notin Y\}$.

Example 6. If $n \geq 2$, \mathbb{Z}_n is an additive abelian group with zero $\bar{0}$ and the negative of \bar{a} being $-\bar{a} = \overline{-a}$. We write $\bar{a} = a$ in \mathbb{Z}_n when no confusion can result.

Henceforth, when we refer to one of the groups \mathbb{Z}_n , \mathbb{Z} , \mathbb{Q} , \mathbb{R} , or \mathbb{C} , we mean the additive group.

Example 7. The set S_n of all permutations of $\{1, 2, \dots, n\}$ is a group under composition (see Theorem 2 §1.4), called the **symmetric group of degree n** .

The group S_n has historical significance because such groups of bijections were among the earliest examples of a group. They were used by Galois in his pioneering work on the theory of equations. In fact Galois was the first to use the term *group*.

Example 8. We single out S_3 for special emphasis. Recall from Section 1.4 that

$$S_3 = \{\varepsilon, (1 \ 2 \ 3), (1 \ 3 \ 2), (1 \ 2), (1 \ 3), (2 \ 3)\}.$$

If we denote $\sigma = (1 \ 2 \ 3)$ and $\tau = (1 \ 2)$, then $\sigma^2 = (1 \ 3 \ 2)$, $\tau\sigma = (2 \ 3)$, and $\tau\sigma^2 = (1 \ 3)$ as is easily verified. Hence we can list S_3 as

$$S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}.$$

The reason for doing this is that it provides an easy way to fill in the Cayley table. In fact, we can fill in the table by using three (easily verified) facts:

$$\sigma^3 = \varepsilon, \quad \tau^2 = \varepsilon, \quad \text{and} \quad \sigma\tau\sigma = \tau.$$

The resulting Cayley table is as follows

S_3	ε	σ	σ^2	τ	$\tau\sigma$	$\tau\sigma^2$
ε	ε	σ	σ^2	τ	$\tau\sigma$	$\tau\sigma^2$
σ	σ	σ^2	ε	$\tau\sigma^2$	τ	$\tau\sigma$
σ^2	σ^2	ε	σ	$\tau\sigma$	$\tau\sigma^2$	τ
τ	τ	$\tau\sigma$	$\tau\sigma^2$	ε	σ	σ^2
$\tau\sigma$	$\tau\sigma$	$\tau\sigma^2$	τ	σ^2	ε	σ
$\tau\sigma^2$	$\tau\sigma^2$	τ	$\tau\sigma$	σ	σ^2	ε

Note that

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

Then, for example, we compute the product $(\tau\sigma)(\tau\sigma^2)$ by

$$(\tau\sigma)(\tau\sigma^2) = \tau(\sigma\tau)\sigma^2 = \tau(\tau\sigma^2)\sigma^2 = \tau^2\sigma^4 = \varepsilon\sigma^4 = \varepsilon\sigma = \sigma.$$

The other entries in the table are found in a similar manner (the reader should do this). The elements σ and τ are called **generators** for S_3 , and the equations $\sigma^3 = \varepsilon$, $\tau^2 = \varepsilon$, and $\sigma\tau\sigma = \tau$ are called **relations** among the generators. We often describe S_3 in this way. \square

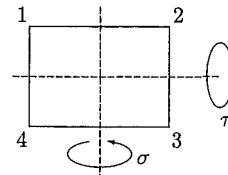
Examples 9 and 10 display two other important groups of permutations.

Example 9. The set A_n of all even permutations in S_n is a group using the operation of S_n , called the **alternating group of degree n** (Theorem 8 §1.4).

Example 10. Given a (nonsquare) wire rectangle with vertices 1, 2, 3, and 4 as in the diagram, consider the permutations of the vertices induced by moving the rectangle in space (without bending). The 180°-rotations about vertical and horizontal axes (see the diagram) give permutations

$$\sigma = (1 \ 2)(3 \ 4) \quad \text{and} \quad \tau = (1 \ 4)(2 \ 3)$$

respectively. If we compute their product in



S_4 we obtain another motion $\sigma\tau = (1 \ 3)(2 \ 4)$ because the composite motion $\sigma\tau$ is the motion τ followed by the motion σ (the reader should verify this). Note that $\sigma\tau$ can also be viewed as the 180°-rotation in the plane of the rectangle about its center. Of course, we have another motion $\tau\sigma$, but this is not a new motion because $\tau\sigma = \sigma\tau$. We do get one more motion $\sigma^2 = \varepsilon$ —no

motion at all. Hence we get a set $K = \{\varepsilon, \sigma, \tau, \sigma\tau\}$ of four motions. This is a group. It is closed because $\sigma^2 = \varepsilon, \tau^2 = \varepsilon$ and $\sigma\tau = \tau\sigma$, and these equations enable us to fill in the entire Cayley table. Since K inherits associativity from S_4 , it is a group because every element is self-inverse. The group K

is called the **group of motions** of the rectangle. Such groups of motions are important (for example they arise in the study of symmetries of molecules); we return to them in Section 2.7. \square

K	ε	σ	τ	$\sigma\tau$
ε	ε	σ	τ	$\sigma\tau$
σ	σ	ε	$\sigma\tau$	τ
τ	τ	$\sigma\tau$	ε	σ
$\sigma\tau$	$\sigma\tau$	τ	σ	ε

Recall that a set M with an associative operation that has a unity is called a monoid, and that an element u in M that has an inverse u^{-1} in M is called a unit. A monoid may not be a group, but its units form a group.

Theorem 1. If M is a monoid, the set M^* of all units in M is a group using the operation of M , called the **group of units** of M .

Proof. From Theorem 5 §2.1, if u and v are units, then uv is also a unit (the inverse is $v^{-1}u^{-1}$), so M^* is closed under the operation of M . The associativity of M^* is inherited from M and $1 \in M^*$ (in fact $1^{-1} = 1$), so M^* itself is a monoid. Finally, if $u \in M^*$, then $u^{-1} \in M^*$ too (its inverse is u), so M^* is a group. \blacksquare

Theorem 1 provides many important examples of groups. For example, the multiplicative groups in Example 2 are $\mathbb{R}^* = \mathbb{R} - \{0\}$, $\mathbb{Q}^* = \mathbb{Q} - \{0\}$, and $\mathbb{C}^* = \mathbb{C} - \{0\}$. Note also that $\mathbb{Z}^* = \{1, -1\}$ and $\mathbb{N}^* = \{1\}$.

Example 11. If X is a nonempty set, $M = \{\alpha \mid \alpha : X \rightarrow X \text{ is a mapping}\}$ is a monoid under composition and Theorem 6 §0.3 shows that the group M^* of units consists of the bijections

$$S_X = \{\alpha \mid \alpha : X \rightarrow X \text{ is a bijection}\}.$$

The bijections $X \rightarrow X$ are called **permutations** of X , and S_X is the **permutation group** of X . Of course, if $X = \{1, 2, \dots, n\}$ then $S_X = S_n$. \square

Example 12. Consider \mathbb{Z}_n^* , where \mathbb{Z}_n is regarded as a multiplicative monoid. Then Theorem 5 §1.3 gives

$$\mathbb{Z}_n^* = \{a \in \mathbb{Z}_n \mid \gcd(a, n) = 1\}.$$

Hence $\mathbb{Z}_p^* = \mathbb{Z}_p - \{0\}$ if (and only if) p is a prime. Other examples include

$$\mathbb{Z}_4^* = \{1, 3\}, \quad \mathbb{Z}_6^* = \{1, 5\}, \quad \mathbb{Z}_8^* = \{1, 3, 5, 7\}, \quad \text{and} \quad \mathbb{Z}_9^* = \{1, 2, 4, 5, 7, 8\}.$$

We refer to these groups frequently. \square

Example 13. Let R denote \mathbb{Z}_m , \mathbb{Z} , \mathbb{Q} , \mathbb{R} , or \mathbb{C} . Then the set $M_n(R)$ of all $n \times n$ matrices over R is a monoid using matrix multiplication. The group $M_n(R)^*$ of units consists of the invertible $n \times n$ matrices over R , that is the matrices such that $\det A$ is a unit in R —see Appendix B. It is called the **general linear group of degree n over R** , denoted $GL_n(R)$. Thus

If $R = \mathbb{Q}$, \mathbb{R} , or \mathbb{C} then $GL_n(R) = \{A \in M_n(R) \mid \det A \neq 0\}$,

$GL_n(\mathbb{Z}) = \{A \in M_n(\mathbb{Z}) \mid \det A = \pm 1\}$, and

$GL_n(\mathbb{Z}_m) = \{A \in M_n(\mathbb{Z}_m) \mid \det A = \bar{a} \text{ where } \gcd(a, m) = 1\}$.

If G_1, G_2, \dots, G_n are sets, recall that the cartesian product $G_1 \times G_2 \times \dots \times G_n$ is the set of all ordered n -tuples (g_1, g_2, \dots, g_n) , where $g_i \in G_i$ for each i . This set has a natural group structure when the G_i are themselves groups. If G_1, G_2, \dots, G_n are groups, their **direct product** is the set $G_1 \times G_2 \times \dots \times G_n$ with the **component-wise operation** defined by

$$(g_1, g_2, \dots, g_n) \cdot (g'_1, g'_2, \dots, g'_n) = (g_1 g'_1, g_2 g'_2, \dots, g_n g'_n)$$

where $g_i g'_i$ is the product in G_i for each i . The routine proof of the next theorem is left to the reader.

Theorem 2. If G_1, G_2, \dots, G_n are groups, so also is $G_1 \times G_2 \times \dots \times G_n$, with unity $(1, 1, \dots, 1)$ and inverses $(g_1, g_2, \dots, g_n)^{-1} = (g_1^{-1}, g_2^{-1}, \dots, g_n^{-1})$.

Because groups are monoids, all the properties of monoids presented in Section 2.1 are automatically properties of groups. In particular:

- (1) The unity 1 is unique.
- (2) The inverse g^{-1} of an element g is uniquely determined by g .
- (3) General associativity holds (Theorem 2 §2.1).

The next theorem restates Theorem 5 §2.1 for units in monoids for reference.

Theorem 3. Let $g, h, g_1, g_2, \dots, g_{n-1}, g_n$ denote elements of a group G .

- (1) $1^{-1} = 1$.
- (2) $(g^{-1})^{-1} = g$.
- (3) $(gh)^{-1} = h^{-1}g^{-1}$.
- (4) $(g_1 g_2 \cdots g_{n-1} g_n)^{-1} = g_n^{-1} g_{n-1}^{-1} \cdots g_2^{-1} g_1^{-1}$ for all $n \geq 1$.
- (5) $(g^n)^{-1} = (g^{-1})^n$ for all $n \geq 0$.

Recall that negative powers of an element g in a group are defined by $g^{-k} = (g^{-1})^k$ for $k \geq 1$. The next theorem is a restatement of Theorem 6 §2.1.

Theorem 4. Exponent Laws. Let G be a group with elements g and h .

- (1) $g^n g^m = g^{n+m}$ for all $n, m \in \mathbb{Z}$.
- (2) $(g^n)^m = g^{n \cdot m}$ for all $n, m \in \mathbb{Z}$.
- (3) If $gh = hg$, then $(gh)^n = g^n h^n$ for all $n \in \mathbb{Z}$.

These laws are important and play a prominent role in Section 2.4.

The assumption that every element of a group has an inverse is a very powerful axiom. In particular, it implies the cancellation laws, which we use countless times in this book.

Theorem 5. Cancellation Laws. Let g, h , and f be elements of a group.

- (1) If $gh = gf$, then $h = f$. (left cancellation)
- (2) If $hg = fg$, then $h = f$. (right cancellation)

Proof. If $gh = gf$, then left multiplication by g^{-1} gives $(g^{-1}g)h = (g^{-1}g)f$. Hence $1h = 1f$; that is, $h = f$. This proves (1), and (2) follows similarly. ■

Note that “mixed” cancellation is not valid in general: $fg = gh$ does *not* imply that $f = h$. For example, in the group S_3 , we have $(1\ 2)(1\ 3) = (1\ 3)(2\ 3)$ so $(1\ 3)$ cannot be cancelled.

Example 14. If G is a finite group and $g \in G$, show that $g^n = 1$ for some $n \geq 1$.

Solution. The elements g, g^2, g^3, \dots in G cannot all be distinct because G is finite. So $g^m = g^{m+n}$ for some $m \geq 1$ and $n \geq 1$. Thus $g^m \cdot 1 = g^m \cdot g^n$, so $1 = g^n$ by cancellation. □

Another consequence of the fact that all elements of a group have inverses is that equations $gx = h$ and $xg = h$ are always solvable.

Theorem 6. Let g and h be elements of a group G .

- (1) The equation $gx = h$ has a unique solution $x = g^{-1}h$ in G .
- (2) The equation $xg = h$ has a unique solution $x = hg^{-1}$ in G .

Proof. If $x = g^{-1}h$, then $gx = gg^{-1}h = 1h = h$, so x is indeed a solution in (1). To prove that it is unique, let y also satisfy $gy = h$. Then $gx = gy$, so $x = y$ by cancellation. This proves (1), and (2) follows in the same way. ■

Corollary. Every row (and column) of the Cayley table of a group G contains every element of G exactly once.

Proof. If $g \in G$, the row of the table corresponding to g consists of the elements gx as x ranges over G . This row contains every element h of G because $gx = h$ is solvable for each h , and it contains h only once because the solution is unique. A similar argument applies to columns. ■

A group is determined completely by its Cayley table: associativity and existence of the unity and inverses, which are demanded by the group axioms, all depend entirely on the operation. Now consider the (multiplicative) group $\mathbb{Z}^* = \{1, -1\}$ of units of \mathbb{Z} and the (additive) group $\mathbb{Z}_2 = \{\bar{0}, \bar{1}\}$. The Cayley tables are

\mathbb{Z}^*	1	-1		\mathbb{Z}_2	0	1
1	1	-1		0	0	1
-1	-1	1		1	1	0

They are the *same* in the sense that the Cayley table of \mathbb{Z}^* becomes that of \mathbb{Z}_2 if we replace the symbols 1 and -1 by $\bar{0}$ and $\bar{1}$, respectively. Thus \mathbb{Z}^* and \mathbb{Z}_2 are the same groups except for notation, and we say that they are **isomorphic**, or that they are the same **up to isomorphism**. We discuss this topic in more detail in Section 2.5; for now we prefer to treat the whole matter informally and call two groups isomorphic if they have the same Cayley table except for notation. As a result we can give an application of the Corollary to Theorem 6.

Example 15. Show that, up to isomorphism, there is only one group G of order 1, 2, or 3, and that group can be described in the following manner.

- If $|G| = 1$, then $G = \{1\}$.
- If $|G| = 2$, then $G = \{1, g\}$, where $g^2 = 1$.
- If $|G| = 3$, then $G = \{1, g, g^2\}$, where $g^3 = 1$.

In each case the Cayley table is determined by the laws of exponents.

Solution. In each case we show that there is only one way to fill in the Cayley table. Multiplication by 1 is prescribed. If $|G| = 1$, then $G = \{1\}$ and the Cayley table is determined. If $|G| = 2$, let $G = \{1, g\}$, where $g \neq 1$. The only entry in the Cayley table that is in doubt is whether $g^2 = g$ or $g^2 = 1$. But $g^2 = g$ is impossible because it implies that $g = 1$ by cancellation. Hence $g^2 = 1$ and the table is determined.

Turning to the case $|G| = 3$, write $G = \{1, g, h\}$. Then $gh \neq g$ and $gh \neq h$ by cancellation, so we must have $gh = 1$. Now repeated use of the Corollary to Theorem 6 (beginning with row 2 and column 3) gives the table on the left.

G	1	g	h	G	1	g	g^2
1	1	g	h	1	1	g	g^2
g	g	h	1	g	g	g^2	1
h	h	1	g	g^2	g^2	1	g

In particular, $g^2 = h$, so $G = \{1, g, g^2\}$, and $g^3 = gh = 1$, as shown in the table on the right. This table is associative, a known realization being the group $\{\varepsilon, (1 \ 2 \ 3), (1 \ 3 \ 2)\}$ of permutations. \square

The groups in Example 15 all have the rather special property that every element is a power of a particular element and are called **cyclic groups**. There exists a cyclic group of order n for every $n \geq 1$. Indeed the group U_n of n th roots of unity is cyclic of order n . In fact, if we write $w = e^{2\pi i/n}$ then $U_n = \{1, w, w^2, \dots, w^{n-1}\}$ has order n and $w^n = 1$.

We discuss cyclic groups in detail in Section 2.4 and treat them informally for now. They occur frequently, and the following generic notation is useful. Given $n \geq 1$, the **cyclic group of order n** is the group C_n of order n :

$$C_n = \{1, a, a^2, \dots, a^{n-1}\}, \quad a^n = 1.$$

We write $C_n = \langle a \rangle$ in this case, and the element a is called a **generator** of C_n . Our insistence that $|C_n| = n$ means that $1, a, a^2, \dots, a^{n-1}$ are distinct elements of C_n .

The Cayley table of C_n is determined completely by the exponent laws and the condition $a^n = 1$. In fact, exponents in C_n can be reduced modulo n . That is, if $k = qn + r$, where $0 \leq r \leq n - 1$, then $a^k = a^r$ because $a^k = (a^n)^q a^r = 1^q a^r = a^r$. In particular,

$$(a^r)^{-1} = a^{n-r} \quad \text{for } r = 0, 1, 2, \dots, n - 1.$$

This expression gives the Cayley table for C_n (below), and so is sufficient for all computations in C_n .

C_n	1	a	a^2	\dots	a^{n-2}	a^{n-1}
1	1	a	a^2	\dots	a^{n-2}	a^{n-1}
a	a	a^2	a^3	\dots	a^{n-1}	1
a^2	a^2	a^3	a^4	\dots	1	a
\vdots	\vdots	\vdots	\vdots	\dots	\vdots	\vdots
a^{n-2}	a^{n-2}	a^{n-1}	1	\dots	a^{n-4}	a^{n-3}
a^{n-1}	a^{n-1}	1	a	\dots	a^{n-3}	a^{n-2}

Example 16. Let $C_{12} = \{1, a, a^2, \dots, a^{11}\}$, $a^{12} = 1$, be a cyclic group of order 12. Compute a^{89} and a^{-40} in C_{12} .

Solution. Because $89 = 7 \cdot 12 + 5$, we get $a^{89} = (a^{12})^7 a^5 = 1^7 a^5 = a^5$. Similarly, $-40 = (-4) \cdot 12 + 8$, so $a^{-40} = (a^{12})^{-4} a^8 = 1^{-4} a^8 = a^8$. \square

Example 17. Show that $b^n = 1$ for every element b of C_n .

Solution. Write $C_n = \langle a \rangle$ where $a^n = 1$. Then $b = a^k$ for some k , so we have

$$b^n = (a^k)^n = a^{kn} = a^{nk} = (a^n)^k = 1^k = 1. \quad \square$$

Example 15 shows that every group of order 1, 2, or 3 is cyclic. However, this is not the case for groups of order 4.

Example 18. Show that there are only two groups of order 4, the cyclic group C_4 and a noncyclic group K_4 whose Cayley table is shown below.

Solution. Let $G = \{1, a, b, c\}$ be any group of order 4.

The way that 1 multiplies is prescribed. Suppose first that $ab = 1$. Then ac cannot be a , 1 or c (by the Corollary to Theorem 6), so $ac = b$. Hence $a^2 = c$, again by the Corollary. In the same way

K_4	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

repeated use of the Corollary shows that the Cayley table is the one on the left below. In that case $a^2 = c$, $a^3 = ca = b$, and $a^4 = c^2 = 1$, so $G = \{1, a, a^2, a^3\} = \langle a \rangle$ is cyclic.

G	1	a	b	c
1	1	a	b	c
a	a	1	b	
b	b	c	a	
c	c	b	a	1

G	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

Similarly if the product of *any two* of a , b , and c equals 1 then G is cyclic (possibly with a different generator). Thus, if G is not cyclic, the product of any two of a , b , and c must equal the third (for example, $bc \neq b, c$, or 1, so $bc = a$). Hence we get the Cayley table on the right as required. \square

The group $K_4 = \{1, a, b, c\}$ in Example 18 is called the **Klein group**.²⁴ The multiplication can be described as follows: $a^2 = b^2 = c^2 = 1$, and the product of any two of a , b , and c is the third.

²⁴The name honors Felix Klein. This group is also called the *four group*.

If you are nervous because we have not shown that K_4 is associative, you can relax. The (associative) group $\mathbb{Z}_8^* = \{1, 3, 5, 7\}$ has exactly the Cayley table of K_4 if we write $a = 3$, $b = 5$, and $c = 7$. Another instance of K_4 is the permutation group $K = \{\epsilon, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$. Example 18 shows that there are two groups of order 4: the cyclic group and the (noncyclic) Klein group. The reader should try to show that every group of order 5 is cyclic; in fact, if p is any prime, we show in Section 2.6 that every group of order p is cyclic.

Exercises 2.2

1. In each case either show that G is a group with the given operation or list the axioms that fail.

- (a) $G = \mathbb{N}$; addition
- (b) $G = \{2n \mid n \in \mathbb{Z}\}$; addition
- (c) $G = \mathbb{R}$; $a \cdot b = a + b + 1$
- (d) $G = \mathbb{R}$; $a \cdot b = a + b - ab$
- (e) $G = \{\epsilon, (1\ 2), (1\ 3), (1\ 4)\}$; operation in S_4
- (f) $G = \{0, 2, 4, 6\}$; addition in \mathbb{Z}_8
- (g) $G = \{16, 12, 8, 4\}$; multiplication in \mathbb{Z}_{20}
- (h) $G = \{q \in \mathbb{Q} \mid q > 0\}$; multiplication
- (i) $G = \{\sigma : \mathbb{N} \rightarrow \mathbb{N} \mid \sigma \text{ is one-to-one}\}$; composition

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

(j) $G = \{a, b, c, d\}$; multiplication given by

2. If G is a group, let G^{op} denote the set G with a new multiplication given by $a \circ b = ba$. Show that G^{op} is a group.

3. In each case fill in the Cayley table, given that $G = \{1, a, b, c, d\}$ is a group.

(a)	G	1	a	b	c	d	(b)	G	1	a	b	c	d
	1	1	a	b	c	d		1	1	a	b	c	d
	a	a	1	b				a	a				
	b	b						b	b		c	d	
	c	c						c	c				
	d	d						d	d				

- 4. Is the empty set a group? Explain.
- 5. If M is a monoid, describe an easy way to determine whether M is a group by looking at the Cayley table.
- 6. If U is a set, let $G = \{X \mid X \subseteq U\}$. Show that G is an abelian group under the operation \oplus defined by $X \oplus Y = (X \setminus Y) \cup (Y \setminus X)$.

- 7. Show that the set $G = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} \mid a, b, c \text{ in } \mathbb{R} \right\}$ is a group under matrix multiplication.
- 8. In each case show that G is a group using the operation of S_4 , and determine how many elements σ of G satisfy $\sigma^2 = \epsilon$.

- (a) $G = \{\varepsilon, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$
 (b) $G = \{\varepsilon, (1\ 2\ 3\ 4), (1\ 3)(2\ 4), (1\ 4\ 3\ 2)\}$
9. Let $\sigma = (1\ 2\ 3\ 4\ 5\ 6)$ in S_6 . Show that $G = \{\varepsilon, \sigma, \sigma^2, \sigma^3, \sigma^4, \sigma^5\}$ is a group using the operation of S_6 . Is G abelian? How many elements τ of G satisfy $\tau^2 = \varepsilon$? $\tau^3 = \varepsilon$?
10. (a) If $a^4 = 1$ and $ab = ba^2$ in a group, show that $a = 1$.
 (b) If $a^6 = 1$ and $ab = ba^3$ in a group, show that $a^2 = 1$ and $ab = ba$.
 (c) If $a^6 = 1$ and $ab = ba^2$ in a group, show that $a^3 = 1$ and $aba = b$.
11. (a) If $(ab)^n = 1$ in a group where $n \geq 0$, show that $(ba)^n = 1$.
 (b) Extend (a) to all $n \in \mathbb{Z}$.
12. Let G be a group of order 4. Assume that 1, a , and b are distinct elements of G and that $a^2 = 1$ and $b^2 = 1$. Show that $G = \{1, a, b, ab\}$ and fill in the Cayley table.
13. If G is any group, define $\alpha : G \rightarrow G$ by $\alpha(g) = g^{-1}$. Show that α is onto and one-to-one.
14. Given a, b , and c in a group G , show that the equation $a^{-1}xb = c$ has a unique solution $x \in G$.
15. Let $a \in G$ where G is a group. If $X \subseteq G$ is a finite subset, write $Xa = \{xa \mid x \in X\}$. Show that X and Xa have the same number of elements.
16. If $fgh = 1$ in a group G , show that $ghf = 1$. Must $gfh = 1$?
17. Recall that an element e in a monoid is called an idempotent if $e^2 = e$. Describe all the idempotents in a group G .
18. If G is a group and $g, h \in G$, show that $gh = hg$ if and only if $g^{-1}h^{-1} = h^{-1}g^{-1}$.
19. Show that a group G is abelian if and only if $(gh)^{-1} = g^{-1}h^{-1}$ for all g and h in G .
20. Show that a group G is abelian if $g^2 = 1$ for all $g \in G$. Give an example showing that the converse is false.
21. Show that a group G is abelian if and only if $(gh)^2 = g^2h^2$ for all g and h in G .
22. Show that a group G is abelian if $(gh)^3 = g^3h^3$, $(gh)^4 = g^4h^4$, and $(gh)^5 = g^5h^5$ for all g and h in G .
23. Let g be an element of a group G .
 (a) Show that $g^2 = 1$ if and only if $g^{-1} = g$.
 (b) If $|G|$ is finite and even, show that $g \neq 1$ in G exists such that $g^2 = 1$.
24. Let a and b be elements of a group G . Prove that $(aba^{-1})^k = ab^k a^{-1}$ holds for all $k \in \mathbb{Z}$ (including negative k).
25. If $a^5 = 1$ and $a^{-1}ba = b^m$ in a group, prove that $b^{m^5-1} = 1$. [Hint: Exercise 24.]
26. Show that every cyclic group C_n of order n is abelian.
27. Show that the additive group \mathbb{Z}_n is cyclic.
28. Let a and b be elements of a group G . If $a^n = b^n$ and $a^m = b^m$ where $\gcd(m, n) = 1$, show that $a = b$. [Hint: Theorem 4 §1.2.]
29. Let G be a set with an associative operation defined on it. In each case show that G is a group.
 (a) There is a left unity e ($eg = g$ for all g in G), and each element g has a left inverse ($hg = e$ for some h in G).
 (b) G is finite and both cancellation laws hold.
 (c) Both $gx = h$ and $xg = h$ are solvable in G for all g and h in G .
 (d) For all g and h in G , $gx = h$ has a unique solution in G .
30. If G is an abelian group with n elements, show that $g^n = 1$ for every $g \in G$. [Hint: See the proof of Theorem 8 §1.3.]

2.3 SUBGROUPS

Many important groups arise as subsets of known groups. Therefore, we are interested in knowing which subsets H of a group G are themselves groups (with the same operation). Thus a subset H of a group G is called a **subgroup** of G if H itself is a group using the operation of G . For example, $(\mathbb{Z}, +)$ is a subgroup of $(\mathbb{R}, +)$. However the multiplicative group (\mathbb{Q}^*, \cdot) is *not* a subgroup of $(\mathbb{R}, +)$, even though \mathbb{Q} is a subset of \mathbb{R} , because the operations are different.

Example 1. If G is any group, both $\{1\}$ and G are subgroups of G . The subgroup $\{1\}$ is the **trivial subgroup** of G . Any subgroup other than G is a **proper subgroup**.

Example 2. Each of the additive groups $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$ is a subgroup of the larger ones.

Example 3. A_n is a subgroup of S_n .

Example 4. $\mathbb{C}^0 = \{z \in \mathbb{C} \mid |z| = 1\}$ denotes the circle group, then each of

$$\{1, -1\} \subseteq \{1, -1, i, -i\} \subseteq \mathbb{C}^0 \subseteq \mathbb{C}^*$$

is a subgroup of the larger ones.

In each of these examples, the subgroups of a group G not only have the same operation as G , but they also share the same unity element and the same inverses. This observation is true in general and, in fact, provides a very useful test for when a subset of a group is actually a subgroup.

Theorem 1. Subgroup Test. A subset H of a group G is a subgroup if and only if the following three conditions are satisfied.

- (1) $1_G \in H$, where 1_G is the unity of G .²⁵
- (2) If $h \in H$ and $h_1 \in H$, then $hh_1 \in H$.
- (3) If $h \in H$ then $h^{-1} \in H$, where h^{-1} denotes the inverse of h in G .

In this case, H has the same unity as G and, if $h \in H$, its inverse in H is the same as its inverse in G .

Proof. If H satisfies (1), (2), and (3), then H is closed by (2), the unity of G is the unity for H by (1), and the inverse in G of an element $h \in H$ serves as the inverse of h in H by (3). As H inherits the associative law from G , it is a subgroup.

Conversely, if H is a subgroup, let e denote the unity of H . Then $e^2 = e = e \cdot 1_G$, so $e = 1_G$ by cancellation in G . This proves (1), and (2) follows because H is closed under the operation of G . Finally, if $h \in H$, let h' denote its inverse in H . If h^{-1} is the inverse in G , then $hh' = 1 = hh^{-1}$, so $h' = h^{-1}$ by cancellation in G . This proves (3) and the last sentence in the theorem. ■

Theorem 1 is useful as the conditions are easily checked (see also Exercise 2).

Example 5. If R is one of \mathbb{Z} , \mathbb{Q} , \mathbb{R} or \mathbb{C} , let $H = \{A \in M_2(\mathbb{R}) \mid \det A = 1\}$. Show that H is a group using matrix multiplication, called the **special linear group**.

²⁵To avoid confusion, we sometimes denote the unity of a group G by 1_G when other groups are present.

Solution. We have $H \subseteq M_2(\mathbb{R})^*$ —see Example 13 §2.2—so we show that it is a subgroup of $M_2(\mathbb{R})^*$. We have $I \in H$ because $\det I = 1$. If A and $B \in H$, then $\det(AB) = \det A \det B = 1 \cdot 1 = 1$ and $\det A^{-1} = 1/\det A = 1/1 = 1$. These results show that $AB \in H$ and $A^{-1} \in H$, so the subgroup test applies. \square

Example 6. If $n \geq 0$, write $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$. Show that $n\mathbb{Z}$ is a subgroup of \mathbb{Z} .

Solution. The unity of \mathbb{Z} is 0, and $0 = n \cdot 0 \in n\mathbb{Z}$. If a and b are in $n\mathbb{Z}$, write them as $a = nk$ and $b = nm$, where $k \in \mathbb{Z}$ and $m \in \mathbb{Z}$. Then $a + b = n(k + m)$ and $-a = n(-k)$ both lie in $n\mathbb{Z}$, so $n\mathbb{Z}$ is a subgroup of \mathbb{Z} by the subgroup test. \square

Theorem 2. Finite Subgroup Test. If H is a finite nonempty subset of a group G , then H is a subgroup of G if and only if H is closed ($h, h_1 \in H$ implies $hh_1 \in H$).

Proof. If H is closed, let $h \in H$. Then each of h, h^2, h^3, \dots is in H so, because H is finite, they cannot all be distinct. Hence $h^n = h^{n+m}$ for some $n \geq 1$ and $m \geq 1$. This means $1 = h^m$ by cancellation, so $1 \in H$ by hypothesis. But then $1 = h^{m-1}h$ implies that $h^{-1} = h^{m-1}$, so $h^{-1} \in H$, too. Because H is closed by hypothesis, it is a subgroup by Theorem 1. The converse is clear. \blacksquare

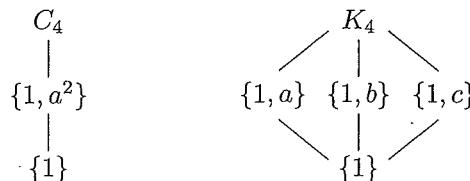
Example 7. Determine all subgroups of the Klein group $K_4 = \{1, a, b, c\}$, where $a^2 = b^2 = c^2 = 1$ and the product of two of a, b , and c is the third.

Solution. Each of $H_a = \{1, a\}$, $H_b = \{1, b\}$, and $H_c = \{1, c\}$ is a subgroup by Theorem 2, because $a^2 = b^2 = c^2 = 1$. Any subgroup H with $|H| \geq 3$ must contain two of a, b , and c and so contains the other one (their product). Thus, $H = G$ and the complete list of subgroups is $\{1\}$, H_a , H_b , H_c , and G . \square

Example 8. Determine all subgroups of $C_4 = \{1, a, a^2, a^3\}$, $a^4 = 1$.

Solution. Let $H = \{1, a^2\}$. Then H is a subgroup by Theorem 2 because $(a^2)^2 = a^4 = 1$. Suppose that K is a subgroup distinct from $\{1\}$ and H . Then either $a \in K$ or $a^3 \in K$. If $a \in K$, then (because K is closed) each power a, a^2 , and a^3 is in K , so $K = C_4$. Similarly, $H = C_4$ if $a^3 \in H$ because, as the reader can verify, $C_4 = \{1, a^3, (a^3)^2, (a^3)^3\}$. Thus the subgroups are $\{1\}$, $H = \{1, a^2\}$, and C_4 . \square

It is descriptive to draw the **lattice diagram** of all subgroups of a group G . Here the subgroups are shown in such a way that a line can be drawn up from K to H whenever $K \subseteq H$. The diagrams for $K_4 = \{1, a, b, c\}$ and for a cyclic group $C_4 = \{1, a, a^2, a^3\}$ of order 4 are given below.



If G is any group, the **center** of G is defined²⁶ by

$$Z(G) = \{z \in G \mid zg = gz \text{ for all } g \in G\}.$$

²⁶The notation $Z(G)$ comes from *zentrum*, the German word for center.

The elements in $Z(G)$ are said to be **central** in G .

Theorem 3. *If G is any group, then $Z(G)$ is an abelian subgroup of G .*

Proof. Use the subgroup test. Clearly $1 \in Z(G)$. If $z \in Z(G)$, then $zg = gz$ for all $g \in G$, so multiplying this equation on the left by z^{-1} gives $g = z^{-1}gz$. Then multiplication on the right by z^{-1} gives $gz^{-1} = z^{-1}g$. Thus $z^{-1} \in Z(G)$. Finally, if both y and z lie in $Z(G)$, then, for all $g \in G$,

$$(yz)g = y(zg) = y(gz) = (yg)z = (gy)z = g(yz).$$

Thus, $yz \in Z(G)$, so $Z(G)$ is a subgroup. It is clearly abelian. ■

Observe that $Z(G) = G$ if and only if G is abelian. At the other extreme, it can happen that $Z(G) = \{1\}$ so G is as far from abelian as it can be. In fact we have:

Example 9. If $n \geq 3$, show that $Z(S_n) = \{\varepsilon\}$, where ε is the identity permutation.

Solution. If $\sigma \in S_n$, $\sigma \neq \varepsilon$, we must find $\tau \in S_n$ such that $\sigma\tau \neq \tau\sigma$. Because $\sigma \neq \varepsilon$, choose k and m in $X_n = \{1, 2, \dots, n\}$ such that $\sigma k = m \neq k$. Because $n \geq 3$, let l, k , and m be distinct, with $l \in X_n$, and take τ to be the transposition $\tau = (k \ l)$. Then $(\tau\sigma)k = \tau m = m$ and $(\sigma\tau)k = \sigma l$, so it suffices to show that $\sigma l \neq m$. But if $\sigma l = m$ then $\sigma l = \sigma k$, so $l = k$ because σ is one-to-one, a contradiction. □

We now turn to two important ways of manufacturing new subgroups from old ones. The straightforward proof of Theorem 4 is left as Exercise 16.

Theorem 4. *Let H and K be subgroups of a group G . then their intersection*

$$H \cap K = \{g \in G \mid g \in H \text{ and } g \in K\}$$

is also a subgroup of G .

Note that $H \cap K$ is a subgroup of both H and K , and is the *largest* such subgroup in the sense that if X is a subgroup of both H and K then $X \subseteq H \cap K$. Incidentally, the union $H \cup K$ of two subgroups is almost never a subgroup (see Exercise 17).

The next theorem introduces another important type of subgroup.

Theorem 5. *Let H be a subgroup of a group G . If $g \in G$, then*

$$gHg^{-1} = \{ghg^{-1} \mid h \in H\}$$

*is a subgroup of G . These subgroups are called the **conjugates** of H in G .*

Proof. Clearly, $1 = g1g^{-1}$ is an element of gHg^{-1} . Given ghg^{-1} , where $h \in H$,

$$(ghg^{-1})^{-1} = (g^{-1})^{-1}h^{-1}g^{-1} = gh^{-1}g^{-1} \in gHg^{-1}.$$

Finally $(ghg^{-1})(gh_1g^{-1}) = g(hh_1)g^{-1}$ for any h, h_1 in H , which shows that gHg^{-1} is closed. Thus it is a subgroup by the subgroup test. ■

If H is a subgroup of G , then $H = 1H1^{-1}$, so H is always a conjugate of itself. If H is the only conjugate of H in G (that is, $gHg^{-1} = H$ for all $g \in G$), then H is said to be **self-conjugate** (or **normal**) in G . These subgroups play a fundamental role in group theory, and will be investigated in detail in Sections 2.8, 2.9, and 2.10. Example 10 displays a subgroup that is not self-conjugate.

Example 10. Let $S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$, where $\sigma^3 = \varepsilon = \tau^2$ and $\sigma\tau\sigma = \tau$. Find the conjugates of the subgroup $H = \{\varepsilon, \tau\}$.

Solution. Clearly $\varepsilon H \varepsilon^{-1} = H$. Since $\sigma^{-1} = \sigma^2$ and $\sigma\tau\sigma = \tau$, we get

$$\sigma H \sigma^{-1} = \{\sigma\varepsilon\sigma^{-1}, \sigma\tau\sigma^{-1}\} = \{\varepsilon, \sigma\tau\sigma^2\} = \{\varepsilon, \tau\sigma\}.$$

Similarly, $\sigma^2 H \sigma^{-2} = \{\varepsilon, \tau\sigma^2\}$. These are all the conjugates of H in G (verify). \square

Exercises 2.3

1. In each case determine whether H is a subgroup of G .
 - (a) $H = \{0, 1, -1\}$, $G = \mathbb{Z}$
 - (b) $H = \{1, 3\}$, $G = \mathbb{Z}_8^*$
 - (c) $H = \{1, 3\}$, $G = \mathbb{Z}_{15}^*$
 - (d) $H = \{\varepsilon, (1 \ 2 \ 3)\}$, $G = S_3$
 - (e) $H = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4)\}$, $G = S_3$
 - (f) $H = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \right\}$, $G = GL_2(\mathbb{Z})$
 - (g) $H = \{2, 4, 6\}$, $G = \mathbb{Z}_6$
 - (h) $H = \mathbb{N}$, $G = \mathbb{Z}$
 - (i) $H = \{(m, k) \mid m + k \text{ is even}\}$, $G = \mathbb{Z} \times \mathbb{Z}$
2. If H is a subset of a group G , show that H is a subgroup if and only if H is nonempty and $ab^{-1} \in H$ whenever $a \in H$ and $b \in H$.
3. If K is a subgroup of H , and H is a subgroup of G , must K be a subgroup of G ? Justify your answer.
4. Let $X = \mathbb{R} \setminus \{0, 1\}$. Show that $G = \{\varepsilon, \lambda_1, \lambda_2, \mu_1, \mu_2, \mu_3\}$ is a subgroup of S_X if $\varepsilon(x) = x$, $\lambda_1(x) = 1/(1-x)$, $\lambda_2(x) = (x-1)/x$, $\mu_1(x) = 1/x$, $\mu_2(x) = x/(x-1)$, and $\mu_3(x) = 1-x$, for all $x \in X$.
5. (a) If G is an abelian group, show that $H = \{a \in G \mid a^2 = 1\}$ is a subgroup of G .
 - (b) Give an example where H is not a subgroup.
6. (a) If G is an abelian group, show that $H = \{g^2 \mid g \in G\}$ is a subgroup of G .
 - (b) Give an example showing that the converse of (a) is false.
 - (c) Show that H is not a subgroup if $G = A_4$.
7. (a) If G is a group and $g \in G$, show that $\langle g \rangle = \{g^k \mid k \in \mathbb{Z}\}$ is a subgroup of G .
 - (b) If G is finite, show that $\{g^k \mid k \in \mathbb{N}\}$ is a subgroup of G for all $g \in G$.
8. If X is a nonempty subset of a group G , let $\langle X \rangle$ be the set of all products of powers of elements of X ; more formally

$$\langle X \rangle = \{x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m} \mid m \geq 1, x_i \in X \text{ and } k_i \in \mathbb{Z} \text{ for each } i\}.$$

- (a) Show that $\langle X \rangle$ is a subgroup of G that contains X .
- (b) Show that $\langle X \rangle \subseteq H$ for every subgroup H such that $X \subseteq H$.

Thus, $\langle X \rangle$ is the *smallest* subgroup of G that contains X , and is called the **subgroup generated by X** .

9. If G is a group and $g \in G$, define $C(g) = \{z \in G \mid zg = gz\}$. Show that $C(g)$ is a subgroup of G (the **centralizer** of g in G).
10. Let $X \subseteq \{1, 2, \dots, n\}$ be a nonempty set. Show that $\{\sigma \in S_n \mid \sigma k = k \text{ for all } k \in X\}$ is a subgroup of S_n .
11. Let $G = \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in \mathbb{R}, a \neq 0 \right\}$. Show that G is a subgroup of $GL_2(\mathbb{R})$.

12. Show that $G = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \mid b \in \mathbb{R} \right\}$ is a subgroup of $GL_2(\mathbb{R})$.
13. (a) If G is a group, show that $\{(g, g) \mid g \in G\}$ is a subgroup of $G \times G$.
(b) Determine the groups G such that $\{(g, g^{-1}) \mid g \in G\}$ is a subgroup of $G \times G$.
14. If X is an infinite set, let G be the set of all permutations σ in S_X such that $\sigma x = x$ for all but a finite number of elements x of X . Show that G is a subgroup of S_X .
15. In each case determine all subgroups of G and draw the lattice diagram.
- (a) $G = C_5$ (b) $G = C_6$ (c) $G = S_3$ (d) $G = \mathbb{Z}_8^*$
16. Let H and K be subgroups of a group G .
(a) Show that $H \cap K$ is a subgroup of G (Theorem 4).
(b) Show that $H \cap K$ is the largest subgroup contained in both H and K in the sense that it contains every subgroup contained in both H and K .
17. If H and K are subgroups of a group G , show that $H \cup K$ is a subgroup if and only if $H \subseteq K$ or $K \subseteq H$.
18. If a and b are real numbers, define $\tau_{a,b} : \mathbb{R} \rightarrow \mathbb{R}$ by $\tau_{a,b}(x) = ax + b$ for all $x \in \mathbb{R}$. Show that $G = \{\tau_{a,b} \mid a, b \in \mathbb{R}, a \neq 0\}$ is a subgroup of $S_{\mathbb{R}}$.
19. Let H and K be subgroups of a group G and let $g \in G$.
(a) If G is abelian, describe the conjugates of H in G .
(b) Show that $(gHg^{-1}) \cap (gKg^{-1}) = g(H \cap K)g^{-1}$.
20. (a) If H is a subgroup of G and $H \subseteq Z(G)$, show that H is self-conjugate in G .
(b) Let $S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$, where $\sigma^3 = \varepsilon = \tau^2$ and $\sigma\tau\sigma = \tau$. Show that $H = \{\varepsilon, \sigma, \sigma^2\}$ is self-conjugate in G .
21. If $G = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \mid a, b, c \in \mathbb{R}, a \neq 0, c \neq 0 \right\}$ find $Z(G)$.
22. Find $Z[GL_2(\mathbb{R})]$.
23. Can a group G have an abelian subgroup not contained in $Z(G)$? Defend your answer.
24. If $ab = ba$ in a group G , let $H = \{g \in G \mid agb = bga\}$. Show that H is a subgroup of G .
25. If H and K are subgroups of G , define $HK = \{hk \mid h \in H, k \in K\}$. Show that HK is a subgroup if and only if $KH \subseteq HK$.

2.4 CYCLIC GROUPS AND THE ORDER OF AN ELEMENT

We have already introduced the cyclic groups C_n , $n \geq 1$, but discussed these groups only informally. Recall that C_n has the form $C_n = \{1, a, \dots, a^{n-1}\}$, where $a^n = 1$, so C_n consists of powers of A . In this section, we classify groups consisting of all powers of a particular element and determine all subgroups of such groups. This endeavor is important because these groups are building blocks for all sufficiently “small” abelian groups (including all finite ones).

We begin by showing that the set of all powers of an element of a group G is an important subgroup of G .

Theorem 1. *Let g be an element of a group G and write*

$$\langle g \rangle = \{g^k \mid k \in \mathbb{Z}\}.$$

Then $\langle g \rangle$ is a subgroup of G , and $\langle g \rangle \subseteq H$ for every subgroup H of G with $g \in H$.

Proof. Clearly, $1 = g^0 \in \langle g \rangle$. If $x, y \in \langle g \rangle$, write them as $x = g^k$, $y = g^m$. Then the exponent laws give $xy = g^{k+m} \in \langle g \rangle$ and $x^{-1} = g^{-k} \in \langle g \rangle$, and the subgroup test applies. Finally, $g = g^1 \in \langle g \rangle$, and if $g \in H$ where H is a subgroup then $\langle g \rangle \subseteq H$ because $g^k \in H$ for all integers k . \blacksquare

Hence, if $g \in G$ then $\langle g \rangle$ is the *smallest* subgroup of G containing the element g .

If g is an element of a group G , the subgroup $\langle g \rangle = \{g^k \mid k \in \mathbb{Z}\}$ is called the **cyclic subgroup of G generated by g** . If $G = \langle g \rangle$ for some $g \in G$, we say that G is a **cyclic group** and that g is a **generator** of G . Thus, the generic cyclic group $C_n = \{1, a, \dots, a^{n-1}\}$, $a^n = 1$, is cyclic in the present sense, so the terminology is consistent.

Example 1. If G is any group, $\{1\} = \langle 1 \rangle$ is a cyclic subgroup of G .

Example 2. The group $G = \{1, -1, i, -i\}$ is cyclic. In fact, $i^2 = -1$ and $i^3 = -i$ show that $G = \langle i \rangle$. Similarly, $G = \langle -i \rangle$, so both i and $-i$ are generators. But -1 is not a generator, because all positive and negative powers of -1 are either 1 or -1 . Hence $\langle -1 \rangle = \{1, -1\}$ is not all of G .

If a group X is written additively, recall that the unity element is denoted 0 and the inverse of $x \in X$ is denoted $-x$. The exponent x^n (in multiplicative notation) becomes nx here, so the cyclic subgroup generated by x is

$$\langle x \rangle = \{kx \mid k \in \mathbb{Z}\} = \mathbb{Z}x$$

consisting of the *multiples* of x . The laws of exponents translate as follows:

$$x^{n+m} = x^n x^m \text{ becomes } (n+m)x = nx + mx,$$

$$(x^n)^m = x^{nm} \text{ becomes } m(nx) = (mn)x,$$

and if x and y commute

$$(xy)^n = x^n y^n \text{ becomes } n(x+y) = nx + ny.$$

Here are two important examples of cyclic additive groups.

Example 3. Show that $(\mathbb{Z}, +)$ is cyclic and that 1 and -1 are the only generators.

Solution. If $k \in \mathbb{Z}$ then $k = k \cdot 1 \in \langle 1 \rangle$, so $\mathbb{Z} = \langle 1 \rangle$. Similarly, $\mathbb{Z} = \langle -1 \rangle$ because $k = (-k) \cdot (-1)$. Clearly $n\mathbb{Z} \neq \mathbb{Z}$, if $n \neq 1$ and $n \neq -1$ (for example $n+1 \notin n\mathbb{Z}$). \square

Example 4. Show that $(\mathbb{Z}_n, +)$ is cyclic with generator $\bar{1}$.

Solution. We have $\mathbb{Z}_n = \{\bar{0}, \bar{1}, \bar{2}, \dots, \bar{n-1}\}$ where, for the moment, we revert to the formal \bar{k} notation for residue classes. Given \bar{k} in \mathbb{Z}_n , note that $\bar{k} = k\bar{1}$ is a multiple of $\bar{1}$, and so $\bar{k} \in \langle \bar{1} \rangle$. It follows that $\mathbb{Z}_n = \langle \bar{1} \rangle$, as required. \square

Example 5. In the multiplicative group \mathbb{R}^* of nonzero real numbers, the cyclic group $\langle 3 \rangle = \{\dots, \frac{1}{27}, \frac{1}{9}, \frac{1}{3}, 1, 3, 9, 27, 81, \dots\}$ consists of all the powers (positive, zero, and negative) of 3 . Note that these powers are all distinct in this case.

Example 6. Consider the group $\mathbb{Z}_7^* = \{1, 2, 3, 4, 5, 6\}$. Here are the powers of 2 :

...	2^{-5}	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^0	2^1	2^2	2^3	2^4	2^5	2^6	...
...	2	4	1	2	4	1	2	4	1	2	4	1	...

If the elements in the bottom row are read left to right they “cycle” endlessly through the sequence 1, 2, 4 (this is the source of the term *cyclic* group). Clearly $\langle 2 \rangle = \{1, 2, 4\}$, and the reason that $\langle 2 \rangle$ has three elements is that 3 is the *smallest* positive integer n such that $2^n = 1$ in \mathbb{Z}_7^* .

Order of an Element

These examples point to one of the most useful concepts in group theory. Let g be an element of a group, and suppose that $g^k = 1$ for some integer $k \neq 0$. Since $g^{-k} = 1$ also holds, we may assume that $k \geq 1$, so the well-ordering principle guarantees that there is a *smallest* integer $n \geq 1$ such that $g^n = 1$. This integer n is called the **order** of g , and is denoted $o(g) = n$. If no such integer n exists we say that g has **infinite order** and write $o(g) = \infty$. To sum up:

1. If $g^k = 1$ for some $k \neq 0$ then $o(g) = n$ is the smallest integer such that $n \geq 1$ and $g^n = 1$.
2. If $g^k = 1$ only if $k = 0$ then $o(g) = \infty$.

In particular, in Example 5, $o(3) = \infty$ in \mathbb{R}^* , while in Example 6, $o(2) = 3$ in \mathbb{Z}_7^* . Note that the unity element 1 is the only element of order 1 in any group.

Example 7. Find the order of each element in $\mathbb{Z}_8^* = \{1, 3, 5, 7\}$. Is \mathbb{Z}_8^* cyclic?

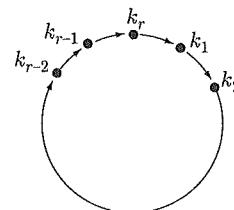
Solution. We have $o(1) = 1$. Since $3^2 = 9 = 1$ in \mathbb{Z}_8^* , it follows $o(3) = 2$. Similarly, $o(5) = 2$ and $o(7) = 2$. Hence no element of \mathbb{Z}_8^* has order 4, so \mathbb{Z}_8^* is not cyclic. \square

Example 8. Find $o \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$ in $GL_2(\mathbb{Z})$.

Solution. Write $A = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$. Then $A^2 = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}$ and $A^3 = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = -I$, so $A^6 = I$. Since $A^4 \neq I$ and $A^5 \neq I$, we conclude that $o \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix} = 6$. \square

Example 9. If $\gamma = (k_1 k_2 \cdots k_r)$ is a cycle in S_n , then $o(\gamma) = r$ is the length of γ .

Solution. If the integers k_1, k_2, \dots, k_r are uniformly placed on a circle, the cycle γ moves each integer one position clockwise, as shown in the diagram. Hence $\gamma^2, \gamma^3, \dots$ carry each integer 2, 3, ... positions clockwise, respectively, so $\gamma^n \neq \varepsilon$ for $1 \leq n \leq r-1$, whereas $\gamma^r = \varepsilon$. This means that $o(\gamma) = r$. \square



Example 10. Show that $o(g^{-1}) = o(g)$ for any group element g .

Solution. If $k \in \mathbb{Z}$ then $(g^{-1})^k = (g^k)^{-1}$, and it follows that $(g^{-1})^k = 1$ if and only if $g^k = 1$. Hence the smallest positive integer n (if any) such that $g^n = 1$ is the same as the smallest positive integer n such that $(g^{-1})^n = 1$. That is, $o(g^{-1}) = o(g)$. \square

Example 11. If G is a finite group, show that every element $g \in G$ has finite order.

Solution. Since G is finite, the powers g, g^2, g^3, \dots are not all distinct, so let $g^k = g^m$ with $k < m$. Then $g^{m-k} = 1$ where $m - k > 0$ so $O(g)$ is finite. \square

Computing the order of an element is simplified by the next theorem

Theorem 2. Let G be a group and let $g \in G$ satisfy $O(g) = n$. Then

- (1) $g^k = 1$ if and only if $n|k$.
- (2) $g^k = g^m$ if and only if $k \equiv m \pmod{n}$.
- (3) $\langle g \rangle = \{1, g, g^2, \dots, g^{n-1}\}$ where $1, g, g^2, \dots, g^{n-1}$ are all distinct.

Proof. We use the laws of exponents.

(1) If $n|k$, say $k = qn$, then $g^k = (g^n)^q = 1^q = 1$. Conversely, if $g^k = 1$, write $k = qn + r$ with $0 \leq r < n$ (division algorithm). But then we have $g^r = g^k(g^n)^{-q} = 1(1)^{-q} = 1$. Since $r < n$, this contradicts the minimality of n unless $r = 0$. So $r = 0$ and $n|k$.

(2) We have $g^k = g^m$ if and only if $g^{k-m} = 1$. Now apply (1).

(3) Clearly, $\{1, g, g^2, \dots, g^{n-1}\} \subseteq \langle g \rangle$. To prove the other inclusion, let $x \in \langle g \rangle$, say $x = g^k$. As before, write $k = qn + r$, where $0 \leq r \leq n - 1$. Then

$$x = g^k = (g^n)^q g^r = 1^q g^r = g^r \in \{1, g, g^2, \dots, g^{n-1}\},$$

which shows that $\langle g \rangle \subseteq \{1, g, g^2, \dots, g^{n-1}\}$. Hence $\langle g \rangle = \{1, g, g^2, \dots, g^{n-1}\}$. To complete the proof, suppose two of $1, g, g^2, \dots, g^{n-1}$ are equal, say $g^k = g^m$, where $0 \leq k \leq m < n$. Then $g^{m-k} = 1$ and $0 \leq m - k < n$. This implies that $m - k = 0$ by the minimality of n , so $g^m = g^k$. Thus $1, g, g^2, \dots, g^{n-1}$ are distinct. \blacksquare

Theorem 2 asserts that if $O(g) = n$, then $g^k = 1$ if and only if $n|k$. The following example illustrates how useful this is.

Example 12. Find the order of 2 in \mathbb{Z}_{19}^* .

Solution. We compute in \mathbb{Z}_{19} : $2^3 = 8$, so $2^6 = 64 = 7$ and $2^9 = 56 = -1$. Hence $2^{18} = 1$, so $O(2)$ divides 18 by Theorem 2. Thus, $O(2)$ is 1, 2, 3, 6, 9, or 18. We have already eliminated 3, 6, and 9, so as $2^1 = 2$ and $2^2 = 4$, the only possibility remaining is $O(2) = 18$. Note that, since $|\mathbb{Z}_{19}^*| = 18$, this shows that \mathbb{Z}_{19}^* is cyclic and that 2 is a generator. \square

The next result is the “companion” of Theorem 2 for elements g with $O(g) = \infty$.

Theorem 3. Let G be a group and let $g \in G$ satisfy $O(g) = \infty$. Then

- (1) $g^k = 1$ if and only if $k = 0$.
- (2) $g^k = g^m$ if and only if $k = m$.
- (3) $\langle g \rangle = \{\dots, g^{-2}, g^{-1}, 1, g, g^2, \dots\}$, where the g^i are distinct.

Proof. (1) Clearly $g^0 = 1$. If $g^k = 1$, $k \neq 0$, then $g^{-k} = (g^k)^{-1} = 1^{-1} = 1$, too. Hence $g^n = 1$ for some $n > 0$, which implies that $\langle g \rangle$ is finite, contrary to hypothesis. Thus $g^k = 1$ implies that $k = 0$.

(2) We have $g^k = g^m$ if and only if $g^{k-m} = 1$. Apply (1).

(3) $\langle g \rangle = \{g^k \mid k \in \mathbb{Z}\}$ by definition, and these powers are distinct by (2). \blacksquare

If $o(g) = n$, then $|\langle g \rangle| = n$ too, by (3) of Theorem 2, so $o(g) = |\langle g \rangle|$ in this case. Since this also holds if $o(g) = \infty$, we have shown that our two uses of the word “order” are compatible.

Corollary. We have $o(g) = |\langle g \rangle|$ for every element g of any group.

We now use Theorem 2 to derive an elegant formula for the order of any permutation σ in S_n . Recall that σ factors (uniquely) as a product of disjoint cycles γ_i (Theorem 5 §1.4). The order of σ turns out to be the least common multiple of the orders of the cycles γ_i (which are the lengths of the γ_i by Example 9).

Theorem 4. Let σ be a permutation in S_t with factorization $\sigma = \gamma_1 \gamma_2 \cdots \gamma_r$ into disjoint cycles. Then $|\sigma| = \text{lcm}(o(\gamma_1), o(\gamma_2), \dots, o(\gamma_r))$.

Proof. Write $n = o(\sigma)$, $n_i = o(\gamma_i)$, and $m = \text{lcm}(n_1, n_2, \dots, n_r)$. As $n_i|m$ for each i , we have $\gamma_i^m = \varepsilon$, and so $\sigma^m = \gamma_1^m \gamma_2^m \cdots \gamma_r^m = \varepsilon$ (because the γ_i commute). Hence $n|m$ by Theorem 2. To show that $m|n$, it suffices to show that $\gamma_i^n = \varepsilon$ for each i (then $n_i|n$ by Theorem 2 so $m|n$ by the definition of the least common multiple). We show that $\gamma_1^n = \varepsilon$; the others are similar. This requires proving that $\gamma_1^n k = k$ for all $0 \leq k \leq n$. This is clear if k is fixed by γ_1 , so let k be moved by γ_1 . Then k is fixed by each of $\gamma_2, \dots, \gamma_r$, because the γ_i are disjoint. Thus, since $\varepsilon = \sigma^n = \gamma_1^n \gamma_2^n \cdots \gamma_r^n$, we have

$$k = \varepsilon k = (\gamma_1^n \gamma_2^n \cdots \gamma_r^n)k = \gamma_1^n (\gamma_2^n \cdots \gamma_r^n)k = \gamma_1^n k.$$

It follows that $\gamma_1^n = \varepsilon$, as required. ■

Example 13. Find the order of

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 5 & 7 & 9 & 14 & 10 & 11 & 12 & 8 & 3 & 13 & 2 & 6 & 4 & 1 \end{pmatrix}.$$

Solution. Here $\sigma = (1 \ 5 \ 10 \ 13 \ 4 \ 14)(2 \ 7 \ 12 \ 6 \ 11)(3 \ 9)$ is the cycle factorization, so Theorem 4 gives $o(\sigma) = \text{lcm}(6, 5, 2) = 30$. □

The next result will be used several times below.

Theorem 5. Let $o(g) = n$ for g in some group. If $d|n$, $d \geq 1$, show that $o(g^d) = \frac{n}{d}$.

Proof. Write $\frac{n}{d} = k$ for convenience. Then $(g^d)^k = g^n = 1$, so we must show that k is the smallest such positive integer. Suppose $(g^d)^r = 1$, $r \geq 1$. Then $g^{dr} = 1$ so $n|dr$ by Theorem 2, say $dr = qn$, $q \geq 1$. But then $dr = q(dk)$, so $r = qk$ because these are integers and $d \neq 0$. It follows that $r \geq k$, as required. ■

Other Properties of Cyclic Groups

Theorem 6. Every cyclic group is abelian, but the converse does not hold.

Proof. Let $G = \langle g \rangle$ be cyclic with generator g . If $x, y \in G$, write $x = g^k$, $y = g^m$, where $k, m \in \mathbb{Z}$. Then the exponent laws give

$$xy = g^k g^m = g^{k+m} = g^{m+k} = g^m g^k = yx,$$

so G is abelian. However \mathbb{Z}_8^* is abelian but not cyclic by Example 7. ■

As the proof of Theorem 6 illustrates, computations in a cyclic group depend entirely on the exponents of the generator. As these exponents are integers, the facts about \mathbb{Z} derived in Chapter 1 turn out to be useful. In particular, the division algorithm plays a natural role in the proof of Theorem 7.

Theorem 7. Every subgroup of a cyclic group is cyclic.

Proof. Suppose that $G = \langle g \rangle = \{g^k \mid k \in \mathbb{Z}\}$ is cyclic and let H be a subgroup of G . If $H = \{1\}$, then $H = \langle 1 \rangle$ is cyclic. Otherwise, let $g^k \in H$, $k \neq 0$. Because H is a subgroup, $g^{-k} = (g^k)^{-1} \in H$, and so we may assume that $k > 0$. Hence let m be the smallest positive integer such that $g^m \in H$. Then $\langle g^m \rangle \subseteq H$, and we claim this is equality. To see this, let $g^k \in H$ and write $k = qm + r$, $0 \leq r < m$, by the division algorithm. It suffices to show that $r = 0$ (then $g^k = (g^m)^q \in \langle g^m \rangle$). But $g^r = (g^m)^{-q}g^k \in H$, which contradicts the minimality of m unless $r = 0$. ■

A cyclic group $G = \langle g \rangle$ can have other generators, for example, $G = \langle g^{-1} \rangle$. Theorem 8 explicitly describes all generators of a finite cyclic group.

Theorem 8. Let $G = \langle g \rangle$ be a cyclic group, where $o(g) = n$. Then $G = \langle g^k \rangle$ if and only if $\gcd(k, n) = 1$.

Proof. If $G = \langle g^k \rangle$, then $g \in \langle g^k \rangle$, say $g = (g^k)^m$, where $m \in \mathbb{Z}$. Thus $g^1 = g^{km}$, so n divides $1 - km$ by Theorem 2. Then $1 - km = qn$ for $q \in \mathbb{Z}$; that is, $1 = km + qn$, which implies that $\gcd(k, n) = 1$. Conversely, if $\gcd(k, n) = 1$ then $1 = xk + yn$ for some integers x and y by Theorem 4 §1.2. Hence

$$g = g^1 = (g^k)^x \cdot (g^n)^y = (g^k)^x \cdot (1)^y = (g^k)^x \in \langle g^k \rangle,$$

which implies that $G = \langle g^k \rangle$. ■

Hence, for example, if $o(g) = 12$ the generators of $G = \langle g \rangle$ are the powers g^k where $\gcd(k, 12) = 1$, that is g, g^5, g^7 , and g^{11} . In particular, the generators of the additive cyclic group \mathbb{Z}_{12} are the residues $\bar{1}, \bar{5}, \bar{7}$, and $\bar{11}$.

Theorem 9 below gives a complete description of all subgroups of a finite cyclic group G . In particular, it shows that G has a unique subgroup of order k for every divisor k of n , and that these are the only subgroups of G .

Theorem 9. Fundamental Theorem of Finite Cyclic Groups. Let $G = \langle g \rangle$ be a cyclic group of order n .

- (1) If H is a subgroup of G , then $H = \langle g^d \rangle$ for some $d|n$. Hence $|H|$ divides n .
- (2) Conversely, if $k|n$, then $\langle g^{n/k} \rangle$ is the unique subgroup of G of order k .

Proof. (1) Theorem 7 implies that $H = \langle g^m \rangle$ for some m . Let $d = \gcd(m, n)$; we show that $H = \langle g^d \rangle$. We have $d|m$, say $m = qd$, so $g^m = (g^d)^q \in \langle g^d \rangle$, when $H \subseteq \langle g^d \rangle$. On the other hand, $d = xm + yn$, for some $x, y \in \mathbb{Z}$, so

$$g^d = (g^m)^x \cdot (g^n)^y = (g^m)^x(1)^y = (g^m)^x \in \langle g^m \rangle = H.$$

Hence $\langle g^d \rangle \subseteq H$, so $\langle g^d \rangle = H$. But then, $|H| = \frac{n}{d}$ by Theorem 5, so $|H|$ divides n .

(2) Suppose that K is any subgroup of G of order k where $k|n$. By (1) let $K = \langle g^d \rangle$ where $d|n$. Then Theorems 2 and 5 give $k = |K| = o(g^d) = \frac{n}{d}$. It follows that $d = \frac{n}{k}$, so $K = \langle g^{n/k} \rangle$. This proves (2). \blacksquare

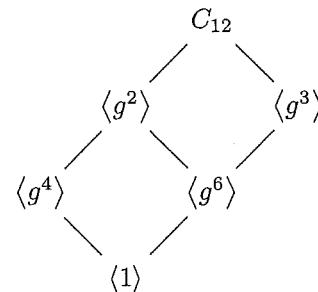
If G is finite and cyclic and H is a subgroup, part (1) of Theorem 9 shows that $|H|$ divides $|G|$. In fact, this result is true for *any* finite group G , cyclic or not. The general result is called Lagrange's theorem, which we prove in Section 2.6.

Example 14. Find all subgroups of C_{12} and draw the lattice diagram.

Solution. Let $C_{12} = \langle g \rangle$, $o(g) = 12$. The divisors of 12 are 1, 2, 3, 4, 6, and 12. Using Theorem 5, the unique subgroup of each of these orders is, respectively,

$$\{1\} = \langle g^{12} \rangle, \langle g^6 \rangle, \langle g^4 \rangle, \langle g^3 \rangle, \langle g^2 \rangle, \\ \text{and } \langle g \rangle = G.$$

The lattice diagram is as shown at the right.
Note that $\langle g^m \rangle \subseteq \langle g^k \rangle$ if and only if $k|m$. \square



We speak of the cyclic subgroup $G = \langle g \rangle$ as being *generated* by the single element g . We conclude this section with a brief discussion of subgroups generated by more than one element.

Theorem 10. Let X be a nonempty subset of a group G and let

$$\langle X \rangle = \{x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m} \mid x_i \in X, k_i \in \mathbb{Z}, m \geq 1\}$$

denote the set of all products of powers of (not necessarily distinct) elements of X . Then

- (1) $\langle X \rangle$ is a subgroup of G containing X .
- (2) If H is a subgroup of G with $X \subseteq H$, then $\langle X \rangle \subseteq H$.

Proof. (1) Choose $x \in X$ (because $X \neq \emptyset$). Then $1 = x^0 \in \langle X \rangle$. The set $\langle X \rangle$ is clearly closed and, if $g = x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m}$ is in $\langle X \rangle$, then $g^{-1} = x_m^{-k_m} \cdots x_2^{-k_2} x_1^{-k_1}$ is also in $\langle X \rangle$. Hence $\langle X \rangle$ is a subgroup of G by the subgroup test.

(2) If $X \subseteq H$ and $g = x_1^{k_1} x_2^{k_2} \cdots x_m^{k_m}$ is in $\langle X \rangle$, then each $x_i^{k_i}$ is in H because $x_i \in X \subseteq H$ and H is a subgroup. Hence $g \in H$, proving (2). \blacksquare

Thus, if X is a nonempty subset of a group G , the subgroup $\langle X \rangle$ in Theorem 10 is the *smallest* subgroup of G that contains X (in the sense of (2) of Theorem 10). Hence $\langle X \rangle$ is called the subgroup **generated** by X . If G has the form $G = \langle X \rangle$ for some $X \subseteq G$, we call X a **set of generators** for G ; if X is finite, we say that G is a **finitely generated group**.

Obviously, $\langle g \rangle = \langle \{g\} \rangle$, so the cyclic groups are exactly the subgroups generated by singleton subsets. Similarly, it is customary to write

$$\langle \{g_1, g_2, \dots, g_n\} \rangle = \langle g_1, g_2, \dots, g_n \rangle$$

for finitely generated groups.

Example 15. Consider the symmetric group $S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$, where $|\sigma| = 3$, $|\tau| = 2$, and $\sigma\tau = \tau\sigma^2$. Then $S_3 = \langle \sigma, \tau \rangle$.

Example 16. The Klein group $K_4 = \{1, a, b, ab\}$ is generated by any two nonunity elements.

Exercises 2.4

1. Find all generators of the cyclic group $G = \langle g \rangle$ if
 - (a) $o(g) = 5$
 - (b) $o(g) = 10$
 - (c) $o(g) = 16$
 - (d) $o(g) = 20$
2. Find all generators of
 - (a) \mathbb{Z}_5
 - (b) \mathbb{Z}_{10}
 - (c) \mathbb{Z}_{16}
 - (d) \mathbb{Z}_{20}
3. Find all generators of
 - (a) $G = \langle g \rangle$, where $o(g) = \infty$
 - (b) \mathbb{Z}
4. In each case determine whether G is cyclic.
 - (a) $G = \mathbb{Z}_7^*$
 - (b) $G = \mathbb{Z}_{12}^*$
 - (c) $G = \mathbb{Z}_{16}^*$
 - (d) $G = \mathbb{Z}_{11}^*$
5. (a) Is \mathbb{Q}^* cyclic? Justify your answer.
 (b) Is \mathbb{Q} cyclic? Justify your answer.
6. If G is a group and $g \in G$, show that $\langle g \rangle = \langle g^{-1} \rangle$.
7. Let $o(g) = 20$ in a group G . Compute
 - (a) $o(g^2)$
 - (b) $o(g^8)$
 - (c) $o(g^5)$
 - (d) $o(g^3)$
8. (a) Find an element of maximum order in S_5 .
 (b) Find an element of maximum order in S_7 .
9. In each case find all subgroups of $G = \langle g \rangle$ and draw the lattice diagram.
 - (a) $o(g) = 8$
 - (b) $o(g) = 10$
 - (c) $o(g) = 18$
 - (d) $o(g) = p^3$, p is a prime.
 - (e) $o(g) = pq$, p and q are distinct primes.
 - (f) $o(g) = p^2q$, p and q are distinct primes.
10. (a) If $gh = hg$ in a group and $o(g)$ and $o(h)$ are finite, show that $o(gh)$ is finite.
 (b) Show that (a) fails if $gh \neq hg$ by considering $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$.
11. Let G be a cyclic group of order n .
 - (a) Show that $g^n = 1$ for all $g \in G$.
 - (b) If $g^m = 1$ in G where $\gcd(m, n) = 1$, show that $g = 1$.
12. Let $g = e^{\frac{2\pi i}{n}}$ in U_n . Show that $o(g) = n$.
13. (a) If $G = \{g_1, g_2, \dots, g_r\}$ is an abelian group, let $a = g_1g_2 \cdots g_r$. Show that $a^2 = 1$.
 (b) Prove **Wilsons Theorem**: $(p-1)! \equiv -1 \pmod{p}$ if p is a prime. [Hint: \mathbb{Z}_p^* .]
14. Suppose that G is a group in which $\{1\}$ and G are the only subgroups. Show that G is finite and, in fact, is cyclic of order 1 or a prime.
15. Show that $\langle a, b \rangle = \langle a, ab \rangle = \langle a^{-1}, b^{-1} \rangle$ for all a and b in a group G .
16. In each case, find the subgroup $H = \langle x, y \rangle$ of G .
 - (a) $G = \langle a \rangle$ is cyclic, $x = a^4$, $y = a^3$
 - (b) $G = \langle a \rangle$ is cyclic, $x = a^6$, $y = a^8$
 - (c) $G = \langle a \rangle$ is cyclic, $x = a^m$, $y = a^k$, $\gcd(m, k) = d$
 - (d) $G = S_3$, $x = (1 \ 2)$, $y = (2 \ 3)$
 - (e) $G = \langle a \rangle \times \langle b \rangle$, $o(a) = 4 = o(b)$, $x = (a^3, b)$, $y = (a, b)$
 - (f) $G = \langle a \rangle \times \langle b \rangle$, $o(a) = 4$, $o(b) = 6$, $x = (a^2, b)$, $y = (a, b^3)$
17. (a) If $X \subseteq Y$ in a group, show that $\langle X \rangle \subseteq \langle Y \rangle$.
 (b) Show that a nonempty subset X is a subgroup if and only if $\langle X \rangle = X$.
18. If $G = \langle g \rangle$ and $H = \langle h \rangle$, show that $G \times H = \langle (g, 1), (1, h) \rangle$.
19. If $G = \langle X \rangle$ and $xy = yx$ for all $x, y \in X$, show that G is abelian.

20. (a) Find three elements of $C_6 \times C_{15}$ of maximum order.
 (b) Find one element of maximum order in $C_m \times C_n$.
21. Find the smallest positive integer n such that $\sigma^n = \varepsilon$ for every $\sigma \in S_5$.
22. If $\sigma \in S_n$ and $o(\sigma) = p$ is a prime, show that σ is a product of disjoint p -cycles.
23. (a) Show that $o(h) = o(ghg^{-1})$ for all $g, h \in G$. [Hint: Example 10.]
 (b) Show that $o(gh) = o(hg)$ for all $g, h \in G$. [Hint: Example 10.]
24. (a) If h is the only element of order 2 in a group G , show that $h \in Z(G)$. [Hint: Exercise 23(a).]
 (b) If a is the unique element of order 3 in G , what can you say about a ?
25. Let G and H be cyclic groups, with $|G| = m$ and $|H| = n$. If $\gcd(m, n) = 1$, show that $G \times H$ is cyclic. [Hint: If $G = \langle g \rangle$ and $H = \langle h \rangle$, use Theorem 5 §1.2 to show $o((g, h)) = mn$.]
26. Let $o(g) = m$ and $o(h) = n$ in a group G , where m and n are relatively prime.
 (a) If $gh = hg$, show that $o(gh) = mn$. Is $o(gh) = \text{lcm}(m, n)$ in general? [Hint: Theorem 5 §1.2.]
 (b) If $o(a) = mn$, show that $a = gh = hg$ for some $g, h \in G$ with $o(g) = m$ and $o(h) = n$. [Hint: Theorem 4 §1.2.]
27. Let $G = \langle g \rangle$ be a cyclic group and let $A = \langle g^a \rangle$ and $B = \langle g^b \rangle$ be cyclic subgroups.
 (a) If $o(g) = \infty$, show that $A \subseteq B$ if and only if $a = qb$ for some $q \in \mathbb{Z}$.
 (b) If $o(g) = n$, show that $A \subseteq B$ if and only if $a \equiv qb \pmod{n}$ for some $q \in \mathbb{Z}$.
28. Let H be a subgroup of a group G and let $a \in G$, $o(a) = n$. If m is the smallest positive integer such that $a^m \in H$, show that $m|n$.
29. If $o(g) = n$, show that $o(g^k) = n/d$, where $d = \gcd(n, k)$. [Hint: Proof of Theorem 9.]
30. Let $G = \langle g \rangle$ where $o(g) = n$. Given $g^k \in G$, show $\langle g^k \rangle = \langle g^d \rangle$, where $d = \gcd(k, n)$. [Hint: Theorem 3 §1.2.]
31. Let $G = \langle g \rangle$ be a cyclic group and let $A = \langle g^a \rangle$ and $B = \langle g^b \rangle$.
 (a) If $o(g) = \infty$, show that $A \cap B = \langle g^m \rangle$, where $m = \text{lcm}(a, b)$.
 (b) If $o(g) = n$, assume (Theorem 9) that $a|n$ and $b|n$. Show that $A \cap B = \langle g^m \rangle$, where $m = \text{lcm}(a, b)$.
32. Show that the following conditions are equivalent for a finite group G .
 (1) G is cyclic and $|G| = p^n$, where p is a prime and $n \geq 0$.
 (2) If H and K are subgroups of G , either $H \subseteq K$ or $K \subseteq H$.
 [Hint: For (1) \Rightarrow (2) use Theorem 8.]
33. If a group G has a finite number of subgroups, show that G must be finite.
34. Prove the **Chinese Remainder Theorem**. Let n_1, n_2, \dots, n_r be positive integers, relatively prime in pairs. Given integers m_1, m_2, \dots, m_r , show that there exists $m \in \mathbb{Z}$ such that $m_i \equiv m \pmod{n_i}$ for each i . [Hint: Extend Exercise 25 to r groups.]
35. (a) Let $o(a) = m$ and $o(b) = n$ in a group G . If $ab = ba$, show that an element $c \in G$ exists, with $o(c) = \text{lcm}(a, b)$. [Hint: Theorem 10 §1.2, Theorem 8, and Exercise 26(a).]
 (b) Let G be an abelian group and assume that G has an element of maximal order n (always true if G is finite). Show that $g^n = 1$ for all $g \in G$. [Hint: Part (a).]
36. Let m be the smallest positive integer such that $\sigma^m = \varepsilon$ for all $\sigma \in S_n$. Show that $m = \text{lcm}(2, 3, 4, 5, \dots, n)$.
37. For a deck of $2n$ distinct cards, a “perfect shuffle” means cutting the deck into two equal halves and collating them as follows: If the cards were originally in the order $1, 2, 3, 4, \dots, 2n$, they end up in the order $1, n+1, 2, n+2, \dots, n, 2n$. In each case, determine the number of perfect shuffles required to bring the deck back into its original order.

- | | |
|---------------------------|-------------------------------|
| (a) $n = 4, 5, 6$, and 7 | (b) $n = 8, 9$, and 10 |
| (c) $n = 12$ | (d) $n = 26$ (a regular deck) |

2.5 HOMOMORPHISMS AND ISOMORPHISMS

Mathematicians do not deal in objects, but in relations among objects; they are free to replace some objects by others so long as the relations remain unchanged. Content to them is irrelevant: they are interested in form only.

—Henri Poincaré

Up to this point we have ignored mappings from one group to another. The interesting ones are those that *preserve* the group multiplication in the following sense: If G and H are groups, a mapping $\alpha : G \rightarrow H$ is called a **homomorphism**²⁷ if

$$\alpha(ab) = \alpha(a) \cdot \alpha(b) \quad \text{for all } a \text{ and } b \text{ in } G.$$

Of course the product ab here is in G while $\alpha(a) \cdot \alpha(b)$ is in H .

Homomorphisms arise in many forms as the following examples illustrate.

Example 1. The mapping $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}$ given by $\alpha(a) = 3a$ is a homomorphism of additive groups because $\alpha(a+b) = 3(a+b) = 3a + 3b = \alpha(a) + \alpha(b)$ for all $a, b \in \mathbb{Z}$.

Example 2. If a is an element of a group G , define the **exponent map** $\alpha : \mathbb{Z} \rightarrow \langle a \rangle$ by $\alpha(k) = a^k$ for all $k \in \mathbb{Z}$. Then α is a homomorphism because (as the operation in \mathbb{Z} is addition)

$$\alpha(k+m) = a^{k+m} = a^k a^m = \alpha(k) \cdot \alpha(m) \quad \text{for all } k, m \in \mathbb{Z}.$$

Example 3. Let \mathbb{R}^+ denote the group of positive real numbers under multiplication. The absolute value map $\alpha : \mathbb{C}^* \rightarrow \mathbb{R}^+$ given by $\alpha(z) = |z|$ for all $z \in \mathbb{C}^*$ is a homomorphism (in fact, onto) because $|zw| = |z||w|$ for all $z, w \in \mathbb{C}$.

Example 4. Let $GL_n(\mathbb{R})$ denote the general linear group of $n \times n$ invertible matrices over \mathbb{R} . The determinant map $GL_n(\mathbb{R}) \rightarrow \mathbb{R}^*$ given by $A \mapsto \det A$ is a homomorphism because $\det(AB) = \det A \det B$ for all matrices A and B , and $\det A \neq 0$ if A is invertible.

Example 5. The identity map $1_G : G \rightarrow G$ is a homomorphism for any group G because $1_G(ab) = ab = 1_G(a) \cdot 1_G(b)$ for all a, b in G .

Example 6. For groups G and H , there is at least one homomorphism from G to H , the **trivial homomorphism** $\alpha : G \rightarrow H$ defined by $\alpha(g) = 1$ for all $g \in G$.

Example 7. If $\alpha : G \rightarrow H$ and $\beta : H \rightarrow K$ are homomorphisms, show that the composite map $\beta\alpha : G \rightarrow K$ is also a homomorphism.

Solution. This is because, for all a and b in G ,

$$\beta\alpha(ab) = \beta[\alpha(ab)] = \beta[\alpha(a) \cdot \alpha(b)] = \beta[\alpha(a)] \cdot \beta[\alpha(b)] = \beta\alpha(a) \cdot \beta\alpha(b). \quad \square$$

By definition, a homomorphism $\alpha : G \rightarrow H$ is a mapping that preserves the operation in the sense that $\alpha(ab) = \alpha(a)\alpha(b)$ for all a and b in G . Theorem 1 shows that α is “structure preserving” in the sense that it also preserves the unity, inverses, and powers.

²⁷Homomorphisms were first used explicitly (for permutation groups) by Jordan in 1870.

Theorem 1. Let $\alpha : G \rightarrow H$ be a group homomorphism. Then

- | | |
|--|---------------------------------|
| (1) $\alpha(1_G) = 1_H$. | (α preserves the unity) |
| (2) $\alpha(g^{-1}) = \alpha(g)^{-1}$ for all $g \in G$. | (α preserves inverses) |
| (3) $\alpha(g^k) = \alpha(g)^k$ for all $g \in G$ and $k \in \mathbb{Z}$. | (α preserves powers) |

Proof. (1) Here $\alpha(1_G) \cdot \alpha(1_G) = \alpha(1_G^2) = \alpha(1_G) = \alpha(1_G) \cdot 1_H$. Now cancel in H .

(2) From (1), $\alpha(g^{-1}) \cdot \alpha(g) = \alpha(g^{-1}g) = \alpha(1) = 1$, which gives (2).

(3) For $k = 0$, $\alpha(g^0) = \alpha(1) = 1 = [\alpha(g)]^0$. If (3) holds for some $k \geq 0$, then

$$\alpha(g^{k+1}) = \alpha(gg^k) = \alpha(g) \cdot \alpha(g^k) = \alpha(g) \cdot [\alpha(g)]^k = [\alpha(g)]^{k+1}.$$

Hence (3) holds for all $k \geq 0$ by induction. If $k < 0$, write $k = -m$, $m > 0$. Then (2) and the preceding calculation give

$$\alpha(g^k) = \alpha[(g^m)^{-1}] = [\alpha(g^m)]^{-1} = [\alpha(g)^m]^{-1} = [\alpha(g)]^{-m} = [\alpha(g)]^k.$$

Thus $[\alpha(g)]^k = \alpha(g^k)$ for all $k \in \mathbb{Z}$. ■

Corollary 1. Let $\alpha : G \rightarrow H$ be a homomorphism. If $g \in G$ has finite order, then $\alpha(g)$ also has finite order, and $o(\alpha(g))$ divides $o(g)$.

Proof. If $o(g) = n$ then $g^n = 1$, so $\alpha(g)^n = \alpha(g^n) = \alpha(1) = 1$. Hence $o(\alpha(g))$ divides n by Theorem 2 §2.4. ■

Corollary 2. If $\alpha : G \rightarrow H$ is a homomorphism, write $\alpha(G) = \{\alpha(g) \mid g \in G\}$. Then $\alpha(G)$ is a subgroup of H .

Proof. This follows from the subgroup test because of the following observations: $1_H = \alpha(1_G) \in \alpha(G)$; $\alpha(g)\alpha(g_1) = \alpha(gg_1) \in \alpha(G)$, and $\alpha(g)^{-1} = \alpha(g^{-1}) \in \alpha(G)$. ■

The group $\alpha(G)$ in Corollary 2 is called the **image** of α . Note that $\alpha : G \rightarrow H$ is onto if and only if $\alpha(G) = H$.

Example 8. Let $\alpha : G \rightarrow H$ be an onto homomorphism.

- (1) If G is abelian show that H is abelian.
- (2) If $G = \langle a \rangle$ is cyclic show that H is cyclic and $H = \langle \sigma(a) \rangle$.

Solution. Let $h, h_1 \in H$. Since α is onto, write $h = \alpha(g)$ and $h_1 = \alpha(g_1)$, $g, g_1 \in G$.

- (1) If G is abelian: $hh_1 = \alpha(g)\alpha(g_1) = \alpha(gg_1) = \alpha(g_1g) = \alpha(g_1)\alpha(g) = h_1h$.
- (2) Let $G = \langle a \rangle$. If $h \in H$, say $h = \alpha(g)$, let $g = a^k$, $k \in \mathbb{Z}$. It suffices to prove that $h \in \langle \alpha(a) \rangle$. But $h = \alpha(a^k) = \alpha(a)^k \in \langle \alpha(a) \rangle$, as required. □

Let G and H denote groups. In order to show that two mappings $\alpha : G \rightarrow H$ and $\beta : G \rightarrow H$ are equal, we must verify that $\alpha(g) = \beta(g)$ holds for all $g \in G$. However, if α and β are homomorphisms, this need only be checked for all g in some generating set for G —see Theorem 10 §2.4.

Theorem 2. Let $\alpha : G \rightarrow H$ and $\beta : G \rightarrow H$ be homomorphisms and assume that $G = \langle X \rangle$ is generated by a subset X . Then

$$\alpha = \beta \quad \text{if and only if} \quad \alpha(x) = \beta(x) \quad \text{for all } x \in X.$$

Proof. If $\alpha = \beta$, the condition is obvious. If the condition holds, let $g \in G$ and write (Theorem 10 §2.4) $g = x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$, where $x_i \in X$ and $k_i \in \mathbb{Z}$ for each i . Then Theorem 1 gives

$$\alpha(g) = \alpha(x_1)^{k_1} \alpha(x_2)^{k_2} \cdots \alpha(x_n)^{k_n} = \beta(x_1)^{k_1} \beta(x_2)^{k_2} \cdots \beta(x_n)^{k_n} = \beta(g).$$

As $g \in G$ was arbitrary, this shows that $\alpha = \beta$. ■

Theorem 2 shows that a group homomorphism $\alpha : G \rightarrow H$ is completely determined by its effect on a generating set for G . This is useful because many groups are generated by a relatively small number of elements.

Example 9. Show that there are at most six homomorphisms $S_3 \rightarrow C_6$.

Solution. As in Example 8 §2.2 we write $S_3 = \{1, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$ where $o(\sigma) = 3$, $o(\tau) = 2$, and $\sigma\tau\sigma = \tau$, and write $C_6 = \langle c \rangle$, $o(c) = 6$. Hence $S_3 = \langle \sigma, \tau \rangle$, so Theorem 2 shows that a homomorphism $\alpha : S_3 \rightarrow C_6$ is determined by the choice of $\alpha(\sigma)$ and $\alpha(\tau)$ in C_6 . Now $\alpha(\sigma)^3 = \alpha(\sigma^3) = \alpha(\varepsilon) = 1$, so $o(\alpha(\sigma))$ is 1 or 3. Hence there are three choices for $\alpha(\sigma)$: 1, c^2 , or c^4 . Similarly, $\alpha(\tau)^2 = 1$, so $\alpha(\tau)$ must be 1 or c^3 . Thus, there are at most $3 \cdot 2 = 6$ choices in all for $\alpha(\sigma)$. □

We hasten to note that *not* all the choices in Example 9 correspond to actual homomorphisms. In fact, there are *only two* homomorphisms from S_3 to C_6 , and we return to this example later (see Example 9 §2.10).

Isomorphisms

We have shown that there are two *distinct* groups of order 4: the cyclic group and the noncyclic Klein group. Determining how to distinguish between distinct groups leads to the notion of isomorphic groups. Roughly speaking, the two groups are isomorphic if they are the same except for notation.

As an illustration, consider the groups $G = \{1, -1\}$ and $\mathbb{Z}_4^* = \{\bar{1}, \bar{3}\}$. The two Cayley tables are

G	1	-1	\mathbb{Z}_4^*	1	3
1	1	-1	1	1	3
-1	-1	1	3	3	1

Clearly, they are alike. In fact, because the way the unity multiplies is always specified, we can describe both by saying that the nonunity element squares to 1. Here is a more precise comparison: Consider the mapping $\sigma : G \rightarrow \mathbb{Z}_4^*$ given by

$$\sigma(1) = 1 \quad \text{and} \quad \sigma(-1) = 3.$$

Then σ is a bijection, and we can obtain the entire Cayley table for \mathbb{Z}_4^* from that of G by replacing a with $\sigma(a)$ for every a in G . In other words, the two groups are the same except for notation; we obtain \mathbb{Z}_4^* from G by changing symbols.

This works in general. If G and H are groups and $\sigma : G \rightarrow H$ is a bijection, we ask when the Cayley table for H results from applying σ to every element of the table for G . Looking at the diagram

G	\dots	b	\dots	H	\dots	$\sigma(b)$	\dots
\vdots				\vdots			
a		ab		$\sigma(a)$		$\sigma(ab)$	
\vdots				\vdots			

the required condition is $\sigma(ab) = \sigma(a)\sigma(b)$ for all a and b in G . In other words, σ must be a homomorphism. This leads to the following definition.

If G and H are groups, a mapping

$$\sigma : G \rightarrow H \text{ is called an isomorphism}$$

if σ is a bijection (one-to-one and onto) which is also a homomorphism. When an isomorphism exists from G to H we say that G is **isomorphic** to H and we write

$$G \cong H.$$

Hence, if $\sigma : G \rightarrow H$ is an isomorphism, the group H is just G with the change of notation $g \mapsto \sigma(g)$. As in the preceding illustration, G and H are the same group except for the symbols used. It is useful to think of isomorphic groups as two different realizations of the same (abstract) group.²⁸

Example 10. The set $2\mathbb{Z} = \{2k \mid k \in \mathbb{Z}\}$ of even integers is an additive group, in fact a subgroup of \mathbb{Z} . Show that $\mathbb{Z} \cong 2\mathbb{Z}$.

Solution. The function $\sigma : \mathbb{Z} \rightarrow 2\mathbb{Z}$ given by $\sigma(k) = 2k$ for all $k \in \mathbb{Z}$ is clearly onto, and σ is one-to-one because $\sigma(k) = \sigma(m)$ implies $2k = 2m$, so $k = m$. Finally, σ is a homomorphism:

$$\sigma(k + m) = 2(k + m) = 2k + 2m = \sigma(k) + \sigma(m)$$

for all k and m in \mathbb{Z} . Thus σ is an isomorphism, so $\mathbb{Z} \cong 2\mathbb{Z}$. \square

Note that the argument in Example 10 shows that $\mathbb{Z} \cong n\mathbb{Z}$ for any integer $n \neq 0$.

Example 11. If $G = \left\{ \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} \mid n \in \mathbb{Z} \right\}$, show that G is a group using matrix multiplication, and that $G \cong (\mathbb{Z}, +)$.

Solution. Define $\sigma : \mathbb{Z} \rightarrow GL_2(\mathbb{R})$ by $\sigma(n) = \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix}$ for all n in \mathbb{Z} . Then σ is clearly one-to-one, and it is a homomorphism because

$$\sigma(m + n) = \begin{bmatrix} 1 & m+n \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & n \\ 0 & 1 \end{bmatrix} = \sigma(m) \cdot \sigma(n).$$

Hence $G = \sigma(\mathbb{Z})$ is a subgroup of $GL_2(\mathbb{R})$ by Corollary 2 of Theorem 1. Moreover, σ is a bijection $\mathbb{Z} \rightarrow G = \sigma(\mathbb{Z})$, so $\sigma : \mathbb{Z} \rightarrow G$ is an isomorphism. \square

Clearly, $G \cong \mathbb{Z}$ for any group G (the identity map $G \rightarrow G$ is an isomorphism). However, even though two groups are isomorphic, they sometimes appear to be quite different. As a remarkable example, the group \mathbb{C}^* of all nonzero complex numbers

²⁸The term *isomorphism* comes from *isos*, meaning *equal*, and *morphe*, meaning *shape*.

is known to be isomorphic to the circle group \mathbb{C}^0 of complex numbers on the unit circle.²⁹ Here is a less spectacular example. Recall that $\mathbb{R}^+ = \{r \in \mathbb{R} \mid r > 0\}$.

Example 12. Show that $\mathbb{R} \cong \mathbb{R}^+$, where \mathbb{R} is additive and \mathbb{R}^+ is multiplicative.

Solution. Define $\sigma : \mathbb{R} \rightarrow \mathbb{R}^+$ by $\sigma(r) = e^r$, where e^x is the exponential function. To show that σ is one-to-one, let $\sigma(r) = \sigma(s)$, where $r, s \in \mathbb{R}$. Then $e^r = e^s$ so, if $\ln x$ denotes the natural logarithm of x , $r = \ln(e^r) = \ln(e^s) = s$. Thus σ is one-to-one. If $t \in \mathbb{R}^+$, then $t > 0$, so $\ln t \in \mathbb{R}$ and $\sigma(\ln t) = e^{\ln t} = t$. Hence σ is onto. Finally,

$$\sigma(r+s) = e^{r+s} = e^r e^s = \sigma(r) \cdot \sigma(s), \quad \text{for all } r \text{ and } s \text{ in } \mathbb{R},$$

which shows that σ is an isomorphism. \square

Example 13. If $G \cong G_1$ and $H \cong H_1$, show that $G \times G_1 \cong H \times H_1$.

Solution. Let $\sigma : G \rightarrow G$ and $\tau : H \rightarrow H_1$ be isomorphisms, and define a mapping $\mu : G \times H \rightarrow G_1 \times H_1$ by $\mu(g, h) = (\sigma(g), \tau(h))$. This is a homomorphism because

$$\begin{aligned} \mu[(g, h)(g', h')] &= \mu(gg', hh') = [\sigma(gg'), \tau(hh')] \\ &= [\sigma(g), \tau(h)][\sigma(g'), \tau(h')] = \mu(g, h) \cdot \mu(g', h'). \end{aligned}$$

for all (g, h) and (g', h') in $G \times H$. The proof that τ is onto and one-to-one is left to the reader. \square

Verifying that a particular mapping is an isomorphism requires checking three things: that it is onto; that it is one-to-one; and that it is operation-preserving. Even if a particular mapping $\alpha : G \rightarrow H$ may fail one of these tests, the groups G and H may very well be isomorphic (for example $r \mapsto r + 1$ is a bijection $\mathbb{R} \rightarrow \mathbb{R}$, but it is not an isomorphism). Conversely, showing that G and H are *not* isomorphic entails showing that *no* isomorphism exists from G to H . Examples 14 and 15 illustrate this situation.

Example 14. Show that \mathbb{Q} is not isomorphic to \mathbb{Q}^* .

Solution. Suppose that $\sigma : \mathbb{Q} \rightarrow \mathbb{Q}^*$ is an isomorphism. Then σ is onto, so let $q \in \mathbb{Q}$ satisfy $\sigma(q) = 2$, and write $\sigma(\frac{1}{2}q) = a$. Since σ is a homomorphism, we have

$$a^2 = \sigma(\frac{1}{2}q) \cdot \sigma(\frac{1}{2}q) = \sigma(\frac{1}{2}q + \frac{1}{2}q) = \sigma(q) = 2.$$

This is impossible because $a \in \mathbb{Q}$ (Example 3 §0.1), so no such σ can exist. \square

Example 15. Let G and H be cyclic groups with $|G| = 9$ and $|H| = 3$. Show that G and $H \times H$ are not isomorphic, even though both groups have order 9.

Solution. Suppose that $\sigma : H \times H \rightarrow G$ is an isomorphism. If $G = \langle a \rangle$ where $o(a) = 9$, let $a = \sigma(x)$ with $x \in H \times H$. Then $x^3 = 1$ (this holds in H) so $a^3 = \sigma(x^3) = 1$, a contradiction. Hence $H \times H \not\cong G$. \square

The reason that $H \times H \not\cong G$ in Example 15 is that, while $x^3 = 1$ for every element x of $H \times H$, this is not the case for G . The condition $x^3 = 1$ for all x is a property of the Cayley table of $H \times H$ but not of the Cayley table of G . The fact that group isomorphisms preserve such properties is the reason that $H \times H$ is not

²⁹See, for instance, Clay, J.R., The punctured plane is isomorphic to the unit circle, *Journal of Number Theory*, 1, (1964), 500–501.

isomorphic to G . More generally, we can often show two groups are not isomorphic by exhibiting such a property that holds in one but not the other.

Theorem 3. Let G , H , and K denote groups.

- (1) The identity map $1_G : G \rightarrow G$ is an isomorphism for every group G .
- (2) If $\sigma : G \rightarrow H$ is an isomorphism, the inverse mapping $\sigma^{-1} : H \rightarrow G$ is also an isomorphism.
- (3) If $\sigma : G \rightarrow H$ and $\tau : H \rightarrow K$ are isomorphisms, the composite map $\tau\sigma : G \rightarrow K$ is also an isomorphism.

Proof. (1) is clear.

(2) The inverse mapping $\sigma^{-1} : H \rightarrow G$ exists because σ is a bijection, and σ^{-1} is also a bijection (Theorems 5 and 6 §0.3). So it remains to show that σ^{-1} is a homomorphism. If g_1 and h_1 are in G_1 , write $g = \sigma^{-1}(g_1)$ and $h = \sigma^{-1}(h_1)$. Then $\sigma(g) = g_1$ and $\sigma(h) = h_1$, so

$$\sigma^{-1}(g_1h_1) = \sigma^{-1}[\sigma(g) \cdot \sigma(h)] = \sigma^{-1}[\sigma(gh)] = gh = \sigma^{-1}(g_1) \cdot \sigma^{-1}(h_1).$$

Therefore σ^{-1} is a homomorphism, and hence is an isomorphism.

- (3) The map $\tau\sigma$ is a bijection by Theorem 3 §0.3; now apply Example 7. ■

Corollary 1. The isomorphic relation \cong is an equivalence for groups. That is

- (1) $G \cong G$ for every group G .
- (2) If $G \cong H$ then $H \cong G$.
- (3) If $G \cong H$ and $H \cong K$ then $G \cong K$.

Proof. (1), (2), and (3) follows from the corresponding items in Theorem 3. ■

As an illustration of Corollary 1, we show that if G and H are both cyclic of order n then $G \cong H$. Indeed $G \cong \mathbb{Z}_n$ and $H \cong \mathbb{Z}_n$ by Example 13, so $G \cong H$ by Corollary 1.

Automorphisms

If G is a group, an isomorphism $G \rightarrow G$ is called an **automorphism** of G .

Corollary 2. If G is a group, the set of all automorphisms $G \rightarrow G$ forms a group under composition.

Proof. The automorphisms $G \rightarrow G$ are a subset of the group S_G of all bijections $G \rightarrow G$, and Theorem 3 shows that they are a subgroup of S_G by the subgroup test. ■

The set of all automorphisms of G is called the **automorphism group** of G , and is denoted $\text{aut } G$.

Example 16. If G is abelian, the mapping $\sigma : G \rightarrow G$ defined by $\sigma(g) = g^{-1}$ for all $g \in G$ is an automorphism of G . We leave the verification as Exercise 10.

If G is a group and $a \in G$, define a mapping $\sigma_a : G \rightarrow G$ by

$$\sigma_a(g) = aga^{-1} \quad \text{for all } g \in G.$$

This map σ_a is an automorphism of G (see Example 17 below), called the **inner automorphism** of G determined by a . Note that if $H \subseteq G$ is a subgroup then $\sigma_a(H) = aHa^{-1}$ is a conjugate of H .

Example 17. If G is any group and $a \in G$, show that

- (1) For each $a \in G$, σ_a is an automorphism of G .
- (2) If $\theta : G \rightarrow \text{aut } G$ is defined by $\theta(a) = \sigma_a$ for each $a \in G$, then θ is a homomorphism, that is $\sigma_{ab} = \sigma_a \sigma_b$ for all $a, b \in G$.
- (3) The image $\theta(G) = \{\sigma_a \mid a \in G\}$ of θ is a subgroup of $\text{aut } G$.

Solution. (1) We leave as Exercise 11 the verification that σ_a is a bijection for all $a \in G$. If $g, h \in G$ we have

$$\sigma_a(g) \cdot \sigma_a(h) = aga^{-1} \cdot aha^{-1} = ag1ha^{-1} = agha^{-1} = \sigma_a(gh).$$

Hence σ_a is an automorphism of G , proving (1).

- (2) We must show that $\sigma_a \sigma_b = \sigma_{ab}$ for $a, b \in G$. But for any $g \in G$:

$$\sigma_a \sigma_b(g) = \sigma_a(bgb^{-1}) = a(bgb^{-1})a^{-1} = (ab)g(ab)^{-1} = \sigma_{ab}(g).$$

- (3) This follows from (2) and Corollary 2 of Theorem 1. □

In Example 17, the group $\theta(G)$ of all **inner automorphisms** of G is denoted

$$\text{inn } G = \{\sigma_a \mid a \in G\}.$$

The group $\text{inn } G$ is an important subgroup of $\text{aut } G$, and it is easily described because each inner automorphism σ_a is given explicitly in terms of a . By contrast, the group $\text{aut } G$ can be difficult to determine. We do one simple case in Example 18 below.

Because it is a homomorphism, every isomorphism preserves the unity, inverses, and powers. But isomorphisms also preserve the order of an element (compare with Corollary 1 of Theorem 1).

Theorem 4. Let $\sigma : G \rightarrow G_1$ be an isomorphism. Then $o(\sigma(g)) = o(g)$ for all $g \in G$.

Proof. It suffices to show that $g^k = 1$ if and only if $[\sigma(g)]^k = 1$. If $g^k = 1$, then $[\sigma(g)]^k = \sigma(g^k) = \sigma(1) = 1$ by Theorem 1. Conversely, if $[\sigma(g)]^k = 1$, we have $\sigma(g^k) = [\sigma(g)]^k = 1 = \sigma(1)$, so $g^k = 1$ because σ is one-to-one. ■

Example 18. If G is cyclic of order 6, show that $\text{aut } G = \{1_G, \lambda\}$, where $\lambda(g) = g^{-1}$ for all $g \in G$.

Solution. Both 1_G and (as G is abelian) λ are automorphisms of G . If $\sigma : G \rightarrow G$ is any automorphism, we show $\sigma = 1_G$ or $\sigma = \lambda$. Write $G = \langle a \rangle$, where $o(a) = 6$. Theorem 2 shows that the choice of $\sigma(a)$ completely determines σ . By Theorem 4, we have $o(\sigma(a)) = o(a) = 6$, so $\sigma(a) = a$, or $\sigma(a) = a^5 = a^{-1}$. If $g \in G$, write $g = a^k$ for some $k \in \mathbb{Z}$, so that

$$\sigma(g) = \sigma(a^k) = [\sigma(a)]^k.$$

If $\sigma(a) = a$, this gives $\sigma(g) = a^k = g$ for all $g \in G$, that is $\sigma = 1_G$. If $\sigma(a) = a^{-1}$, it shows that $\sigma(g) = (a^{-1})^k = (a^k)^{-1} = g^{-1}$ for all $g \in G$, that is $\sigma = \lambda$. □

Cayley's Theorem

We conclude this section with a proof of a theorem of Cayley (proved in 1878) that every finite group is isomorphic to a group of permutations. If X is a nonempty set, recall that S_X denotes the group of all permutations of X (bijections $X \rightarrow X$) under composition. We need one simple observation about these permutation groups: If a bijection $\sigma : X \rightarrow Y$ exists then $S_X \cong S_Y$. Indeed, if $\lambda \in S_X$ we have

$$Y \xrightarrow{\sigma^{-1}} X \xrightarrow{\lambda} X \xrightarrow{\sigma} Y$$

so $\sigma\lambda\sigma^{-1} \in S_Y$. But then $\varphi : S_X \rightarrow S_Y$ given by $\varphi(\lambda) = \sigma\lambda\sigma^{-1}$ is an isomorphism, as can be readily verified. In particular, $S_X \cong S_n$ whenever $|X| = n$.

Now let G be a group. We noted earlier that each row of the Cayley table of G is a permutation of G in the sense that each element appears exactly once. Since the row of $a \in G$ is $\{ag \mid g \in G\}$, this is just the assertion that $g \mapsto ag$ is a bijection $G \rightarrow G$. This is the connection that Cayley noticed between the groups G and S_G .

Theorem 5. Cayley's Theorem. Every group G of order n is isomorphic to a subgroup of S_n .

Proof. By the preceding discussion, there is an isomorphism $\theta : S_G \rightarrow S_n$. So if we can find a one-to-one homomorphism $\sigma : G \rightarrow S_G$, then $G \cong \theta\sigma(G) \subseteq S_n$ because $\theta\sigma : G \rightarrow \theta\sigma(G)$ is an isomorphism, and Cayley's theorem follows.

If $a \in G$, define $\mu_a : G \rightarrow G$ by $\mu_a(g) = ag$ for all $g \in G$. Then it is easy to verify that μ_a is a bijection (so $\mu_a \in S_G$). Hence define $\theta : G \rightarrow S_G$ by $\sigma(a) = \mu_a$ for all $a \in G$. Then θ is a homomorphism because $\mu_{ab} = \mu_a\mu_b$ for all $a, b \in G$ (verify). Finally, θ is one-to-one because $\mu_a = \mu_b$ implies that $a = \mu_a(1) = \mu_b(1) = b$. So σ is a one-to-one homomorphism, as required. ■

Cayley's theorem shows that every abstract group of order n is (up to isomorphism) a subgroup of S_n . Hence, to study the groups of order n , we need only study the symmetric group S_n . At first this approach seems to be an advantage because S_n consists of concrete mappings that can be analyzed using tools (such as cycle factorization and parity) not available in an abstract group. However, these symmetric groups are extremely large, so a subgroup of order n is lost in S_n (for example, $|S_{10}| = 10! = 3,628,800$). However, in Section 8.3 we give a generalization of Cayley's theorem that cuts down the size of the symmetric group and so provides more information about G .

Arthur Cayley (1821–1895) Cayley showed his mathematical talent at an early age, quickly excelling at school. After some initial reluctance, his merchant father sent him to Cambridge at the age of 17. During the following 8 years he read the works of the masters and published more than 20 papers on topics that would occupy him for the rest of his life. In addition, he developed broad interests in literature (he read Greek, German, and French, as well as English), architecture, and painting (he demonstrated talent in watercolors) and became an enthusiastic hiker and mountaineer.

At the age of 25, with no position as a mathematician in view, he began legal training and was admitted to the bar three years later. He earned a comfortable living as a lawyer but resisted the temptation to make a lot of money so as to free himself to

do mathematics. And do it he did, publishing nearly 300 papers in 14 years. Finally, in 1863, he accepted the Sadlerian professorship at Cambridge and remained there for the rest of his life, valued for his administrative and teaching skills, as well as for his scholarship.

Although Cayley introduced the concept of an abstract group, his main accomplishments lay elsewhere. With his lifelong friend J. J. Sylvester, he founded the theory of invariants; he was one of the first to consider geometry of more than three dimensions; and he initiated matrix algebra. He also wrote on quaternions, the theory of equations, dynamics, and astronomy. He continued working until his death, leaving 966 papers filling 13 volumes of 600 pages each.

Exercises 2.5

- In each case show that α is a homomorphism and decide if it is onto or one-to-one.
 - $\alpha : \mathbb{R} \rightarrow GL_2(\mathbb{R})$ given by $\alpha(r) = \begin{bmatrix} 1 & r \\ 0 & 1 \end{bmatrix}$ for all r in \mathbb{R} .
 - $\alpha : G \rightarrow G \times G$ given by $\alpha(g) = (g, g)$ for all g in the group G .
- If $G = G_1 \times G_2$ is a direct product of groups, define $\pi_1 : G \rightarrow G_1$ and $\sigma_1 : G_1 \rightarrow G$ by $\pi_1(g_1, g_2) = g_1$ and $\sigma_1(g_1) = (g_1, 1)$. Show that π_1 is an onto homomorphism (called the *projection* of G onto G_1), and σ_1 is a one-to-one homomorphism (called the *injection* of G_1 into G).
- If G is any group, define $\alpha : G \rightarrow G$ by $\alpha(g) = g^{-1}$. Show that G is abelian if and only if α is a homomorphism.
- If $m \in \mathbb{Z}$ is fixed and G is an abelian group, define $\alpha : G \rightarrow G$ by $\alpha(a) = a^m$ for all $a \in G$. Show that α is a homomorphism.
- Let σ_a be the inner automorphism of G determined by a . Show that $\sigma_a = 1_G$ if and only if $a \in Z(G)$.
- Show that there are exactly two homomorphisms $\alpha : C_6 \rightarrow C_4$. [Hint: Example 9.]
- If $n \geq 1$, give an example of a group homomorphism $\sigma : G \rightarrow G_1$ and an element $g \in G$ such that $\sigma(g) = \infty$ but $\sigma(\alpha(g)) = n$.
- (a) Describe all group homomorphisms $\mathbb{Z} \rightarrow \mathbb{Z}$.
(b) How many are onto?
- If $\alpha : G \rightarrow G_1$ is a homomorphism, show that $K = \{g \in G \mid \alpha(g) = 1\}$ is a subgroup of G (called the *kernel* of α).
- Define $\lambda : G \rightarrow G$ by $\lambda(g) = g^{-1}$ for all $g \in G$. Show that λ is a bijection. If G is abelian, show that λ is an automorphism of G .
- If G is a group and $a \in G$, show that the inner automorphism $\sigma_a : G \rightarrow G$ is a bijection.
- In each case determine whether $\alpha : G \rightarrow G_1$ is an isomorphism. Give reasons.
 - $G = G_1 = \mathbb{R}$, $\alpha(x) = 2x$
 - $G = G_1 = \mathbb{Z}$, $\alpha(n) = 2n$
 - $G = G_1 = \mathbb{Z}_5^*$, $\alpha(g) = g^2$
 - $G = G_1 = \mathbb{Z}_5^*$, $\alpha(g) = g^3$
 - $G = G_1 = \mathbb{Z}_7$, $\alpha(g) = 2g$
 - $G = G_1 = \mathbb{Z}_8$, $\alpha(g) = 2g$
 - $G = G_1 = \mathbb{R}^+$, $\alpha(g) = g^2$
 - $G = G_1 = \mathbb{R}^+$, $\alpha(g) = o(g)$
 - $G = 2\mathbb{Z}$, $G_1 = 3\mathbb{Z}$, $\alpha(2k) = 3k$
 - $G = G_1 = \mathbb{R}$, $\alpha(g) = ag$, $a \neq 0$
- Show that $G = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \right\}$ is a subgroup of $GL_2(\mathbb{Z})$ isomorphic to $\{1, -1, i, -i\}$.
- If G is an infinite cyclic group, show that $G \cong \mathbb{Z}$.

15. If $G = \langle a \rangle$ is cyclic with $o(a) = n$, show that $G \cong \mathbb{Z}_n$. [Hint: $\bar{k} = \bar{m}$ in \mathbb{Z}_n if and only if $a^k = a^m$ by Theorem 2 §2.4.]
16. Show that $\sigma : \mathbb{C}^* \rightarrow \mathbb{C}^*$ is an automorphism if $\sigma(z) = \bar{z}$ for all $z \in \mathbb{C}$ (here \bar{z} denotes the complex conjugate of z).
17. If g and h are elements of a group G , show that $\langle gh \rangle \cong \langle hg \rangle$.
18. If G is a group of order 2, show that $G \times G \cong K_4$.
19. If $\sigma : G \rightarrow G_1$ is an isomorphism, show that $Z(G_1) = \sigma[Z(G)]$, where we have $\sigma[Z(G)] = \{\sigma(z) \mid z \in Z(G)\}$.
20. Write $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$. Show that $n\mathbb{Z} \cong m\mathbb{Z}$ whenever $n \neq 0$ and $m \neq 0$.
21. Show that \mathbb{Z}_{10}^* is not isomorphic to \mathbb{Z}_{12}^* .
22. Show that \mathbb{R} is not isomorphic to \mathbb{R}^* .
23. Show that the circle group $\mathbb{C}^0 = \{z \in \mathbb{C} \mid |z| = 1\}$ is not isomorphic to \mathbb{R}^* .
24. Find two nonisomorphic groups of order n^2 for any integer $n \geq 2$.
25. Are the additive groups \mathbb{Z} and \mathbb{Q} isomorphic? Support your answer.
26. Show that $\mathbb{Z}_{14}^* \cong \mathbb{Z}_{18}^*$.
27. If $G = \langle a \rangle$ and $G_1 = \langle b \rangle$, where $o(a) = o(b) = 6$, describe all isomorphisms $G \rightarrow G_1$.
28. Show that $\mathbb{R}^+ \times \mathbb{C}^0 \cong \mathbb{C}^*$, where $\mathbb{C}^0 = \{z \in \mathbb{C} \mid |z| = 1\}$ is the circle group.
29. Define $\tau_{a,b} : \mathbb{R} \rightarrow \mathbb{R}$ by $\tau_{a,b}(x) = ax + b$ for all $x \in \mathbb{R}$, and denote $G_1 = \{\tau_{a,b} \mid a, b \in \mathbb{R}, a \neq 0\}$. Let $G = \left\{ \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \mid a, b \in \mathbb{R}, a \neq 0 \right\}$. Show that G and G_1 are subgroups of $GL_2(\mathbb{R})$ and $S_{\mathbb{R}}$, respectively, and that $G \cong G_1$.
30. If $G = \left\{ \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \mid a, b \in \mathbb{R}, a \text{ and } b \text{ not both } 0 \right\}$, show that G is a subgroup of $M_2(\mathbb{R})^*$ and that $G \cong \mathbb{C}^*$.
31. In each case, find $\text{aut } G$, where $G = \langle a \rangle$ is cyclic of order n : (a) $n = 2$ (b) $n = 3$
32. If G is infinite cyclic, determine $\text{aut } G$.
33. If $Z(G) = \{1\}$, show that $G \cong \text{inn } G$.
34. Given $z \in Z(G)$, let G^z denote the set G with a new operation $a * b = abz^{-1}$. Show that G^z is a group and $G^z \cong G$.
35. If G is a group and $g \in G$, let $S(g) = \{\sigma \in \text{aut } G \mid \sigma(g) = g\}$.
- (a) Show that $S(g)$ is a subgroup of $\text{aut } G$ for all $g \in G$.
 - (b) If $g_1 = \tau(g)$, $\tau \in \text{aut } G$, show that $S(g)$ and $S(g_1)$ are conjugate in $\text{aut } G$.
36. In a group G , write $a \sim b$ if $b = gag^{-1}$ for some $g \in G$ (a is *conjugate* to b).
- (a) Show that \sim is an equivalence relation on G .
 - (b) Determine which elements of G have singleton equivalence classes.
37. If $G = \langle X \rangle$ and $\sigma : G \rightarrow G_1$ is an onto homomorphism, show that $G_1 = \langle \sigma(X) \rangle$, where $\sigma(X) = \{\sigma(x) \mid x \in X\}$.
38. Show that $\mathbb{Z}_{15}^* \cong \mathbb{Z}_{16}^*$.

2.6 COSETS AND LAGRANGE'S THEOREM

He [Lagrange] would set to mathematics all the little themes on physical inquiries which his friends brought him, much as Schubert set to music any stray rhyme that took his fancy.

—Herbert Westron Turnbull

In this section we prove one of the most important theorems about finite groups, Lagrange's theorem, which asserts that the order of a subgroup of a finite group G is a divisor of $|G|$. This has far-reaching consequences as we shall see. The proof involves counting elements of G , and depends on the following basic notion.

Let H be a subgroup of a group G . If $a \in G$ we identify two subsets of G :

$$\begin{aligned} Ha &= \{ha \mid h \in H\} \text{ — the right coset of } H \text{ generated by } a. \\ aH &= \{ah \mid h \in H\} \text{ — the left coset of } H \text{ generated by } a. \end{aligned}$$

We have $H1 = H = 1H$, so H is a right and left coset of itself. Also the fact that $1 \in H$ shows $a \in Ha$ and $a \in aH$ for all a . Of course, if G is abelian then $Ha = aH$ for all $a \in G$ and all subgroups H of G . However, this may not hold if G is not abelian (see Example 5 below).

Example 1. Let $K_4 = \{1, a, b, ab\}$ be the Klein group where $o(a) = o(b) = 2$ and $ab = ba$. If $H = \{1, a\}$, find the cosets of H in K_4 .

Solution. $H1 = H = \{1, a^2\} = \{1, a\} = H$ too. Similarly, $Hb = \{b, ab\}$ and $Hab = \{ab, a^2b\} = \{ab, b\} = Hb$. Thus, there are exactly two cosets of H in K_4 : $H = \{1, a\}$ and $Hb = \{b, ab\} = bH$. \square

Note that the cosets $H = \{1, a\}$ and $\{b, ab\}$ form a partition³⁰ of K_4 . This holds in general and, with the other properties in Theorem 1, makes finding cosets easier.

Theorem 1. Let H be a subgroup of a group G and let $a, b \in G$.

- (1) $H = H1$.
- (2) $Ha = H$ if and only if $a \in H$.
- (3) $Ha = Hb$ if and only if $ab^{-1} \in H$.
- (4) If $a \in Hb$, then $Ha = Hb$.
- (5) Either $Ha = Hb$ or $Ha \cap Hb = \emptyset$.
- (6) The distinct right cosets of H are the cells of a partition of G .

Proof. First, (1) is clear because $1 \in H$ and (2) follows from (3) with $b = 1$.

(3). If $Ha = Hb$ then $a \in Ha = Hb$, say $a = hb$, $h \in H$. Hence $ab^{-1} = h \in H$. Conversely, suppose that $ab^{-1} \in H$. Then $ha = h(ab^{-1})b \in Hb$, so $Ha \subseteq Hb$. But $ba^{-1} = (ab^{-1})^{-1} \in H$ too, so $Hb \subseteq Ha$ follows in the same way. Hence $Ha = Hb$.

- (4) If $a \in Hb$ then $ab^{-1} \in H$ so $Ha = Hb$ by (3).
- (5) If $Ha \cap Hb \neq \emptyset$, we show $Ha = Hb$. If $x \in Ha \cap Hb$, then $x \in Ha$ so $Hx = Ha$ by (4). Similarly $Hx = Hb$, so $Ha = Hx = Hb$. This proves (5).
- (6) If $Ha \neq Hb$ then Ha and Hb are disjoint by (5). In other words, the set of right cosets is pairwise disjoint. Moreover, each $a \in G$ belongs to *some* right coset of H (in fact $a \in Ha$). This gives (6). \blacksquare

Corollary. The analogue of Theorem 1 for left cosets also holds. In particular, (3) becomes $aH = bH$ if and only if $b^{-1}a \in H$. See Exercise 5.

³⁰Recall (Section 0.4) that a **partition** of a nonempty set X is a collection of nonempty subsets of X (called the **cells** of the partition) which are **pairwise disjoint** (distinct cells are disjoint) and every element of X is in some cell (hence in exactly one cell).

Mnemonic. The condition in (3) that $Ha = Hb$ if and only if $ab^{-1} \in H$ can be remembered by “right multiplying” $Ha = Hb$ by b^{-1} . Similarly, $aH = bH$ if and only if $b^{-1}a \in H$ can be recalled by “left multiplying” by b^{-1} .

Example 2. Let $G = \langle a \rangle$ where $o(a) = 6$. Find the right cosets of the subgroups $H = \langle a^3 \rangle$ and $K = \langle a^2 \rangle$.

Solution. We have $H = H1 = \{1, a^3\}$. Thus $a^3 \in H$ so $H = Ha^3$ by (4) of Theorem 1. In the same way $Ha = \{a, a^4\} = Ha^4$, and $Ha^2 = \{a^2, a^5\} = Ha^5$. This exhausts G so the cosets are

$$H = \{1, a^3\}, \quad Ha = \{a, a^4\}, \quad \text{and} \quad Ha^2 = \{a^2, a^5\}.$$

Turning to K we find the partition in one step:

$$K = \{1, a^2, a^4\} \quad \text{and} \quad Ka = \{a, a^3, a^5\}. \quad \square$$

In Example 2, the cosets of H (and those of K) do indeed partition G into pairwise disjoint cells, as Theorem 1(6) asserts. What is new here is that all the cosets of H have the same number of elements and, similarly, all the cosets of K have the same number of elements. This fact holds in general and lies at the heart of Lagrange’s theorem, as we show shortly.

Example 3. Find all the right cosets of the subgroup $4\mathbb{Z}$ in the additive group \mathbb{Z} .

Solution. The notation is additive, so the right coset of $4\mathbb{Z}$ generated by a is $4\mathbb{Z} + a$. For $a = 0$, we obtain the coset $4\mathbb{Z}$ itself:

$$4\mathbb{Z} = 4\mathbb{Z} + 0 = \{4k \mid k \in \mathbb{Z}\}.$$

Now $1 \notin 4\mathbb{Z}$, so it generates a new coset by Theorem 1:

$$4\mathbb{Z} + 1 = \{4k + 1 \mid k \in \mathbb{Z}\}.$$

We continue in this way, with 2 and 3 generating new cosets:

$$4\mathbb{Z} + 2 = \{4k + 2 \mid k \in \mathbb{Z}\},$$

$$4\mathbb{Z} + 3 = \{4k + 3 \mid k \in \mathbb{Z}\}.$$

This is a complete list of cosets, because every integer has the form $4k + r$, where the remainder r is 0, 1, 2, or 3. \square

Example 4. In the group \mathbb{C}^* give a geometrical description of the cosets of the circle group $\mathbb{C}^0 = \{z \in \mathbb{C} \mid |z| = 1\}$.

Solution. Recall that $\mathbb{C}^0 = \{e^{i\theta} \mid \theta \text{ any angle}\}$ is the unit circle. If $z \in \mathbb{C}^*$, then $z = re^{i\theta}$, where $r = |z| > 0$ and $e^{i\theta} \in \mathbb{C}^0$. Hence $\mathbb{C}^0 z = \mathbb{C}^0 r = \{re^{i\theta} \mid \theta \text{ any angle}\}$. In other words, $\mathbb{C}^0 z$ is the circle with its center at the origin and radius $r = |z|$. \square

All these examples involve abelian groups so $Ha = aH$ always holds. However, this does not hold in general as Example 5 shows.

Example 5. Let $G = S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$, where $\sigma^3 = \varepsilon = \tau^2$ and $\sigma\tau\sigma = \tau$. Find the right and left cosets of $H = \{\varepsilon, \tau\}$.

Solution. As $\sigma\tau = \tau\sigma^{-1} = \tau\sigma^2$, the cosets are

$$H = H\varepsilon = \{\varepsilon, \tau\}, \quad H\sigma = \{\sigma, \tau\sigma\}, \quad H\sigma^2 = \{\sigma^2, \tau\sigma^2\},$$

$$H = \varepsilon H = \{\varepsilon, \tau\}, \quad \sigma H = \{\sigma, \tau\sigma^2\}, \quad \sigma^2 H = \{\sigma^2, \tau\sigma\}.$$

Observe that $H\sigma \neq \sigma H$ and $H\sigma^2 \neq \sigma^2 H$ in this case. \square

Note that, even though the right and left cosets of H may be different, they all have the same number of elements and there are the same number of them. This holds in general, even when the cosets are infinite sets.

Two sets X and Y are said to have the *same cardinality* if there is a bijection from one to the other, and in this case we write $|X| = |Y|$. Of course if the sets are finite this means that they have the same number of elements, and this terminology is sometimes used for infinite sets too.

Lemma. Let H be a subgroup of a group G . Then

- (1) $|H| = |Ha| = |aH|$ for all $a \in G$.
- (2) The map $Ha \mapsto a^{-1}H$ is a bijection $\{Ha \mid a \in G\} \rightarrow \{bH \mid b \in G\}$.

Proof. (1) $|H| = |aH|$ since $h \mapsto ha$ is a bijection $H \rightarrow Ha$. Similarly, $|H| = |Ha|$.

(2) We have $Ha = Hb \Leftrightarrow ab^{-1} \in H \Leftrightarrow a^{-1}H = b^{-1}H$ by Theorem 1 and its Corollary. So the map in (2) is well defined and one-to-one. It is clearly onto. ■

Part (2) of the Lemma shows that the sets of right and left cosets of a subgroup H of G have the same number of members (possibly infinite), and this common value has a name:

The index $|G : H|$ of H in G is defined to be the number of distinct right (or left) cosets of H in G .

Note that a subgroup H can be of finite index in G even if both H and G are infinite (for example $|\mathbb{Z} : 4\mathbb{Z}| = 4$ by Example 3).

The Lemma enables us to prove the single most important theorem about finite groups: It introduces numerical relations into the theory.

Theorem 2. Lagrange's Theorem. Let H be any subgroup of a finite group G .

- (1) Then $|H|$ divides $|G|$.
- (2) The quotient $\frac{|G|}{|H|} = |G : H|$ is the index of H in G .

Proof. Write $k = |G : H|$, and let Ha_1, Ha_2, \dots, Ha_k be the distinct right cosets of H in G . Then

$$G = Ha_1 \cup Ha_2 \cup \dots \cup Ha_k$$

which is a disjoint union by Theorem 1. By the Lemma, $|Ha_i| = |H|$ for each i , so

$$\begin{aligned} |G| &= |Ha_1| + |Ha_2| + \dots + |Ha_k| \\ &= |H| + |H| + \dots + |H| \\ &= k|H|. \end{aligned}$$

This proves (1), and it also proves (2) because $\frac{|G|}{|H|} = k = |G : H|$. ■

Note that Lagrange's theorem shows that both the order *and* the index of a subgroup of a finite group G are divisors of $|G|$.

Lagrange's theorem has many important consequences.

Corollary 1. If G is a finite group and $g \in G$, then $o(g)$ divides $|G|$.

Proof. The cyclic subgroup $H = \langle g \rangle$ generated by g has $o(g) = |H|$ by the Corollary to Theorem 3 §2.4. So Lagrange's theorem applies. ■

Note that the converse of Lagrange's theorem (and of Corollary 1) is false. For example, $|A_4| = 12$, but A_4 has no subgroup of order 6 and hence no element of order 6 (Exercise 34).

Corollary 2. If G is a group and $|G| = n$, then $g^n = 1$ for every $g \in G$.

Proof. If $o(g) = m$ then $m|n$ by Corollary 1, say $n = qm$ for some $q \in \mathbb{Z}$. But then $g^n = (g^m)^q = 1^q = 1$. ■

The next corollary will be referred to later, and illustrates how the numerical information in Lagrange's theorem can determine the structure of a finite group.

Corollary 3. If p is a prime, then every group G of order p is cyclic. In fact, $G = \langle g \rangle$ for every element $g \neq 1$ in G , so the only subgroups of G are $\{1\}$ and G .

Proof. Let $g \neq 1$ in G and write $H = \langle g \rangle$. Then $|H|$ divides $|G| = p$, so $|H| = 1$ or $|H| = p$. But $|H| \neq 1$ because H contains both 1 and $g \neq 1$, so $|H| = p = |G|$. This implies that $H = G$ because G is finite. Finally, if $K \neq \{1\}$ is a subgroup of G , and $1 \neq k \in K$, then $G = \langle k \rangle \subseteq K \subseteq G$, and so $K = G$. ■

Corollary 4. Let H and K be finite subgroups of a group G . If $|H|$ and $|K|$ are relatively prime, then $H \cap K = \{1\}$.

Proof. As $H \cap K$ is a subgroup of both H and K , $|H \cap K|$ must divide both $|H|$ and $|K|$ by Lagrange's theorem. Since $|H|$ and $|K|$ are relatively prime, it follows that $|H \cap K| = 1$. The Corollary follows. ■

Example 6. Let $K \subseteq H \subseteq G$ be finite groups. If $|G : K|$ is a prime, show that $H = K$ or $H = G$.

Solution. By Lagrange's theorem, $|G : H| \cdot |H : K| = \frac{|G|}{|H|} \cdot \frac{|H|}{|K|} = \frac{|G|}{|K|} = |G : K|$. Since $|G : K|$ is a prime, either $|G : H| = 1$ or $|H : K| = 1$; that is either $H = G$ or $H = K$. □

We showed earlier (Example 18 §2.2) that every group of order $4 = 2^2$ is either cyclic or is isomorphic to the Klein group. In Example 7, Lagrange's theorem is used to give an analogous result for any prime p in place of 2.

Example 7. If G is a group and $|G| = p^2$ where p a prime, show that either G is cyclic or $g^p = 1$ for every element $g \in G$.

Solution. Assume that G is not cyclic. Then $o(g) \mid p^2$, so $o(g) = 1, p$, or p^2 . But $o(g) \neq p^2$ because G is not cyclic, so $o(g)$ is 1 or p . Either way $g^p = 1$. □

Dihedral Groups

Recall (Example 8 §2.2) that the group S_3 can be presented as follows:

$$S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}, \quad o(\sigma) = 3, o(\tau) = 2, \text{ and } \sigma\tau\sigma = \tau.$$

In fact, we can take $\sigma = (1 \ 2 \ 3)$ and $\tau = (1 \ 2)$, but the point here is that the three conditions $o(\sigma) = 3$, $o(\tau) = 2$, and $\sigma\tau\sigma = \tau$ are themselves sufficient to fill in the Cayley table of the group.

We now construct a family of groups $D_2, D_3, \dots, D_n, \dots$ each presented in much the same way as S_3 , and having $D_3 \cong S_3$. We realize them as subgroups of the group $GL_2(\mathbb{C})$ of 2×2 invertible matrices with complex entries.

Let $n \geq 2$ be fixed and let $w = e^{2\pi i/n}$ (an n th root of unity). Then $o(w) = n$ in \mathbb{C}^* . Consider the matrices:

$$\mathbf{a} = \begin{bmatrix} w & 0 \\ 0 & w^{-1} \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

in $GL_2(\mathbb{C})$. It is easy to verify that $o(\mathbf{a}) = n$, $o(\mathbf{b}) = 2$, and $\mathbf{aba} = \mathbf{b}$. This last equation shows that $\mathbf{ab} = \mathbf{ba}^{-1} = \mathbf{ba}^{n-1}$, and hence that the finite set

$$G = \{\mathbf{I}, \mathbf{a}, \mathbf{a}^2, \dots, \mathbf{a}^{n-1}, \mathbf{b}, \mathbf{ba}, \mathbf{ba}^2, \dots, \mathbf{ba}^{n-1}\}$$

of matrices is closed under matrix multiplication (\mathbf{I} is the 2×2 identity matrix). Hence G is a subgroup of $GL_2(\mathbb{C})$ by Theorem 2 §2.3. For convenience, write $A = \langle \mathbf{a} \rangle$. Then $|A| = n$ and, as $\mathbf{b} \notin A$, the left cosets A and $\mathbf{b}A$ are disjoint. Hence $|G| = 2n$. We abstract this situation as follows.

If $n \geq 2$, the **dihedral group** D_n is the group of order $2n$ presented as follows:

$$D_n = \{1, a, a^2, \dots, a^{n-1}, b, ba, ba^2, \dots, ba^{n-1}\},$$

where $o(a) = n$, $o(b) = 2$, and $aba = b$.

Note that the requirement that $|D_n| = 2n$ is equivalent to insisting that $b \notin \langle a \rangle$. We can carry out all calculations in D_n by using the conditions $o(a) = n$, $o(b) = 2$, and $aba = b$. The equation $aba = b$ implies that $a^kba^k = b$ for all $k \in \mathbb{Z}$ by induction (Exercise 25), and so we obtain

$$a^k b = ba^{-k} = ba^{n-k} \quad \text{and} \quad o(ba^k) = 2 \quad \text{for all } k \in \mathbb{Z}.$$

In particular, $ab = ba^{n-1}$, and these formulas enable us to fill in the Cayley table for D_n . Hence the conditions $o(a) = n$, $o(b) = 2$, and $aba = b$ completely determine the group D_n (up to isomorphism). The group D_4 is called the **octic group**, and its Cayley table is as follows (the reader should verify this):

D_4	1	a	a^2	a^3	b	ba	ba^2	ba^3
1	1	a	a^2	a^3	b	ba	ba^2	ba^3
a	a	a^2	a^3	1	ba^3	b	ba	ba^2
a^2	a^2	a^3	1	a	ba^2	ba^3	b	ba
a^3	a^3	1	a	a^2	ba	ba^2	ba^3	b
b	b	ba	ba^2	ba^3	1	a	a^2	a^3
ba	ba	ba^2	ba^3	b	a^3	1	a	a^2
ba^2	ba^2	ba^3	b	ba	a^2	a^3	1	a
ba^3	ba^3	b	ba	ba^2	a	a^2	a^3	1

The group D_3 is isomorphic to S_3 because

$$D_3 = \{1, a, a^2, b, ba, ba^2\}, \quad o(a) = 3, o(b) = 2, \text{ and } aba = b$$

which is the same as the presentation of S_3 given previously. If $n = 2$,

$$D_2 = \{1, a, b, ba\}, \quad o(a) = 2, o(b) = 2, \text{ and } aba = b.$$

We have $a^{-1} = a$ here because $o(a) = 2$, so D_2 is abelian ($ba = a^{-1}b = ab$) and is isomorphic to the Klein group K_4 .

Thus, every group of order 4 is either cyclic or dihedral (Example 18 §2.2). The next theorem shows that this result holds for groups of order $2p$ where p is a prime.

Theorem 3. *Let G be a group of order $2p$ where p is a prime. Then either G is cyclic or $G \cong D_p$.*

Proof. First, the theorem is true if $p = 2$ because $|G| = 4$ implies G is cyclic or $G \cong K_4 \cong D_2$. Hence we assume that p is odd.

Assume that G is not cyclic. Hence $o(g) = 1, 2$, or p for every $g \in G$ by Corollary 1 of Lagrange's theorem. We must show that $G \cong D_p$.

Claim 1. G has an element of order p .

Proof. If not, $g^2 = 1$ for all $g \in G$, so G is abelian by Exercise 20 §2.2. Hence if $1, a$, and b are distinct in G , then $\{1, a, b, ab\}$ is a subgroup of order 4 by Theorem 2 §2.3, contrary to Lagrange's theorem. This proves Claim 1.

So let $a \in G$ have order p and write $H = \langle a \rangle = \{1, a, a^2, \dots, a^{p-1}\}$.

Claim 2. If $x \in G$ and $x \notin H$, then $o(x) = 2$.

Proof. We have $G = H \cup Hx$ so, because $x^2 \notin Hx$, we must have $x^2 \in H$. If $o(x) = p$ then, since p is odd, $x = x^{p+1} = (x^2)^{\frac{p+1}{2}} \in H$, contrary to the choice of x . Thus $o(x) \neq p$, so $o(x) = 2$ ($x \neq 1$ because $x \notin H$). This proves Claim 2.

Now choose $b \notin H$. Then $G = H \cup bH$, a disjoint union, so we obtain

$$G = \{1, a, a^2, \dots, a^{p-1}, b, ba, ba^2, \dots, ba^{p-1}\}.$$

As $o(b) = 2$ by Claim 2, it remains to show that $aba = b$. But $ba \notin H$ so $(ba)^2 = 1$ again by Claim 2. But then $aba = b^{-1}(ba)^2 = b^{-1} = b$. Thus $G \cong D_p$. ■

Theorem 3 together with Corollary 3 determines all groups G with $|G| \leq 7$:

$ G $	1	2	3	4	5	6	7
G	$C_1 = \{1\}$	C_2	C_3	C_4, K_4	C_5	C_6, D_3	C_7

Note that $K_4 \cong C_2 \times C_2$, so every abelian group here is (isomorphic to) a direct product of cyclic groups. In fact this is true for *every* finite abelian group, an important result discussed in Chapter 7.

Obviously, the list continues. We will show that there are five nonisomorphic groups of order 8: C_8 , $C_4 \times C_2$, $C_2 \times C_2 \times C_2$, D_4 , and another group Q called the *quaternion group*, to be introduced in Section 2.8. The groups of order 9 are C_9 and $C_3 \times C_3$ —both abelian,) and there are two distinct groups of order 10; C_{10} and D_5 . The next interesting case is the groups of order 12 (there are five). However, it is not our intention to imply that all the distinct groups of order n have been determined for an arbitrary integer n . That is a *very* difficult task!

We conclude with an application of Lagrange's theorem to number theory. If $n \geq 2$, the **Euler function** φ is defined by

$\varphi(n)$ is the number of integers $k \in \{1, 2, \dots, n-1\}$ with $\gcd(k, n) = 1$.

We define $\varphi(1) = 1$. Hence $\varphi(2) = 1$, $\varphi(3) = 2$, $\varphi(4) = 2$, $\varphi(5) = 4$, and $\varphi(6) = 2$. Clearly,

$$\varphi(p) = p - 1 \text{ whenever } p \text{ is a prime.}$$

Now recall (Theorem 5 §1.3) that, for $n \geq 2$, the group of (multiplicative) units in \mathbb{Z}_n is given by $\mathbb{Z}_n^* = \{k \mid 1 \leq k < n \text{ and } \gcd(k, n) = 1\}$. Hence

$$\text{If } n \geq 2 \text{ then } \varphi(n) = |\mathbb{Z}_n^*|.$$

With this, Lagrange's theorem yields an elegant proof of the following famous result in number theory.

Theorem 4. Euler's Theorem. *If a and $n \geq 2$ are relatively prime integers, then $a^{\varphi(n)} \equiv 1 \pmod{n}$.*

Proof. We have $\bar{a} \in \mathbb{Z}_n^*$. Since $|\mathbb{Z}_n^*| = \varphi(n)$, Lagrange's theorem (Corollary 2) gives $\bar{a}^{\varphi(n)} = \bar{1}$ in \mathbb{Z}_n^* . Euler's theorem follows. ■

A special case gives another proof of Fermat's theorem (Theorem 8 §1.3).

Corollary. Fermat's Theorem. *If p is a prime, then $a^p \equiv a \pmod{p}$ for all integers a .*

Proof. This is clear if $a \equiv 0 \pmod{p}$. Otherwise, a and p are relatively prime so, because $\varphi(p) = p - 1$, Euler's theorem gives $a^{p-1} \equiv 1 \pmod{p}$. Fermat's theorem follows. ■

Joseph Louis Lagrange (1736–1813) While his name sounds French, Lagrange was born in Italy and spent his early years in Turin. In 1766, he was appointed as Euler's successor at the Berlin Academy by Frederick the Great, who suggested that the "greatest mathematician in Europe" should be at the court of the "greatest king in Europe." After the death of Frederick, Lagrange went to Paris at the invitation of Louis XVI. He remained there throughout the revolution and was made a count by Napoleon who called him the "lofty pyramid of the mathematical sciences."

Lagrange was one of the great mathematicians of all time. He made important contributions to many parts of mathematics, including number theory, the theory of equations, differential equations, celestial mechanics, and fluid dynamics. At age 19 he solved a famous problem, the so-called isoperimetrical problem, by inventing an entirely new method, known today as the calculus of variations. His work brought a new level of rigor to analysis. In addition to his mathematical achievements, he was a master of exposition, and his *Mécanique Analytique* is a masterpiece that William Rowan Hamilton described as a "scientific poem,".

In his work on the theory of polynomial equations, Lagrange studied the permutations of the roots of an equation in the hope of finding a general method of solution. He saw that, because the symmetric groups S_2 , S_3 , and S_4 were sufficiently "nice" a general solution can always be found if the degree is 2, 3, or 4. But he never discovered what it was about S_5 that obstructed the solution of equations of degree 5. Abel, and later Galois, eventually clarified the matter. Nevertheless, Lagrange's work provided one of the sources from which the modern theory of groups evolved.

Exercises 2.6

1. In each case find the right and left cosets in G of the subgroups H and K of G .
 - (a) $G = \langle a \rangle$, $o(a) = 20$; $H = \langle a^4 \rangle$, $K = \langle a^2 \rangle$
 - (b) $G = A_4$; $H = \{\epsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$, $K = \langle (1 \ 2 \ 3) \rangle$

- (c) $G = \mathbb{Z}$; $H = 2\mathbb{Z}$, $K = 3\mathbb{Z}$
 (d) $G = \mathbb{Z}_{12}$; $H = 3\mathbb{Z}_{12}$, $K = 2\mathbb{Z}_{12}$
 (e) $G = D_4 = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$, $o(a) = 4$, $o(b) = 2$, and $aba = b$;
 $H = \langle a^2 \rangle$, $K = \langle b \rangle$.
 (f) G = any group; H is any subgroup of index 2
2. If G is any group, describe the cosets in G of the subgroups $\{1\}$ and G .
3. If H is a subgroup of G and $Ha = Hb$ where $a, b \in G$, does it follow that $aH = bH$? Support your answer.
4. If $K \subseteq H \subseteq G$ are finite groups, show that $|G : K| = |G : H| \cdot |H : K|$.
5. If H is a subgroup of G and $a, b \in G$, define $a \equiv b$ if $b^{-1}a \in H$.
- (a) Show that \equiv is an equivalence relation on G .
 - (b) Show that the equivalence class (Section 0.4.) of $a \in G$ is the left coset aH .
6. Let $G = \mathbb{R} \times \mathbb{R}$ with addition $(x, y) + (x', y') = (x + x', y + y')$. Let H be the line $y = mx$ through the origin: $H = \{(x, mx) \mid x \in \mathbb{R}\}$. Show that H is a subgroup of G and describe the cosets $H + (a, b)$ geometrically.
7. Let H be a subgroup of G and suppose that $Ha = bH$ for $a, b \in G$. Show that $aH = bH$.
8. Let H and K be subgroups of G . If $Ha \subseteq Kb$ for some $a, b \in G$, show that $H \subseteq K$.
9. In each case give a geometric description of the cosets of H in G .
- (a) $G = \mathbb{R}^*$, $H = \mathbb{R}^+$
 - (b) $G = \mathbb{C}^*$, $H = \mathbb{R}^*$
 - (c) $G = \mathbb{R}$, $H = \mathbb{Z}$
 - (d) $G = \mathbb{C}$, $H = \mathbb{R}$
10. (a) If $G = \langle a \rangle$ and $o(a) = 30$, find the index of $\langle a^6 \rangle$ in G .
 (b) Let $G = \langle a \rangle$, $o(a) = n$. If $d|n$, find the index of $\langle a^d \rangle$ in G .
11. Let H and K be subgroups of some group G .
- (a) Show that $Ha \cap Ka = (H \cap K)a$ for all $a \in G$.
 - (b) Given $a, b \in G$, show that either $Ha \cap Kb$ is empty or $Ha \cap Kb = (H \cap K)c$ for some $c \in G$.
12. Let G denote a group and let $g \in G$. In each case show $G = \langle g \rangle$.
- (a) $|G| = 12$, $g^4 \neq 1$, $g^6 \neq 1$.
 - (b) $|G| = 40$, $g^8 \neq 1$, $g^{20} \neq 1$.
 - (c) $|G| = 60$, $g^{30} \neq 1$, $g^{20} \neq 1$, and $g^{12} \neq 1$.
 - (d) Generalize. [Hint: Prime factorization.]
13. Let $K = \{\epsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$, and let H be a subgroup of A_4 containing K . If H contains any 3-cycle, show that $H = A_4$.
14. Suppose that G has subgroups of orders 45 and 75. If $|G| < 400$, determine $|G|$.
15. If H and K are subgroups of a group and $|H|$ is prime, show that either $H \subseteq K$ or $H \cap K = \{1\}$.
16. Let G be a group of order n and let m be an integer with $\gcd(m, n) = 1$.
- (a) If $g^m = 1$ in G , show that $g = 1$. [Hint: Theorem 4 §1.2.]
 - (b) Show that each $g \in G$ has an m th root, that is that $g = a^m$ for some $a \in G$.
17. Let $|G| = p^2$, where p is a prime. Show that every proper subgroup of G is cyclic.
18. Let $|G| = p^3$, where p is a prime. If G is not cyclic, show that $g^{p^2} = 1$ for all $g \in G$.
19. Let $a^k = b^k$ in a group. If $o(a) = m$ and $o(b) = n$, where n and m are relatively prime, show that mn divides k . [Hint: Lagrange Corollary 4 and Theorem 5 §1.2.]
20. Show that $|\mathbb{Z}_n^*|$ is even if $n \geq 3$. [Hint: Corollary 1 of Lagrange's theorem.]
21. Show that $|\mathbb{Z} : n\mathbb{Z}| = n$ for every $n \geq 1$.
22. If G is a group of order n , define $\sigma : G \rightarrow G$ by $\sigma(g) = g^m$ for all $g \in G$. If $\gcd(m, n) = 1$, show that σ is a bijection (an automorphism if G is abelian).

23. If G is a group of order p^k , where p is a prime and $k \geq 1$, show that G must have an element of order p . [Hint: Theorem 5 §2.4.]
24. If G is a group of order pq , where p and q are primes, show that every proper subgroup of G is cyclic.
25. (a) In D_n , show that $\alpha^kba^k = b$ for all $k \in \mathbb{Z}$.
 (b) In D_n , show that $o(ba^k) = 2$ for all $k \in \mathbb{Z}$.
26. If $n \geq 3$, show that $Z(D_n) = \{1\}$ if n is odd, and that $Z(D_{2m}) = \{1, a^m\}$.
27. Is $D_5 \times C_3 \cong D_3 \times C_5$? Prove your answer.
28. If $k|n$, $k \geq 2$, show that D_n has a subgroup isomorphic to D_k .
29. Let G be a group and let p be a prime.
 (a) If H and K are subgroups of order p , show that $H = K$ or $H \cap K = \{1\}$.
 (b) If H_1, H_2, \dots, H_k are distinct subgroups of order p , show that

$$|H_1 \cup H_2 \cup \dots \cup H_k| = 1 + k(p - 1).$$

- (c) If $|G| = 15$, show that G must have an element of order 3.
30. Let G be any group (possibly infinite) that has no subgroups except $\{1\}$ and G . If $|G| \geq 2$, show that G is finite and cyclic and that $|G|$ is prime. (Converse of Corollary 3 of Lagrange's theorem.)
31. Let $K \subseteq H \subseteq G$ be groups. Show that both $|G : H|$ and $|H : K|$ are finite if and only if $|G : K|$ is finite, and then $|G : K| = |G : H||H : K|$.
 [Hint: If $|H : K| = n$, let Kh_1, Kh_2, \dots, Kh_n be the distinct cosets of K in H . Show that $Hg = Kh_1g \cup Kh_2g \cup \dots \cup Kh_ng$ is a disjoint union for all $g \in G$.]
32. Let H and K be subgroups of a group G with $|G : H| = m$ and $|G : K| = n$.
 (a) Show that $|G : H \cap K| \leq mn$. [Hint: $(H \cap K)g = Hg \cap Kg$ for all $g \in G$.]
 (b) If $\gcd(m, n) = 1$, show that $|G : H \cap K| = mn$. [Hint: Exercise 31.]
33. Prove Poincaré's Theorem: If H_1, H_2, \dots, H_n are subgroups of a group G of finite index, then $H_1 \cap H_2 \cap \dots \cap H_n$ is also of finite index. [Hint: Exercise 32.]
34. Show that A_4 has no subgroup of order 6, and hence that the converse of Lagrange's theorem is false. [Hint: Theorem 3.]
35. If H and K are subgroups of a group G , define a relation \equiv on G by $a \equiv b$ if $a = hbk$ for some $h \in H$ and $k \in K$.
 (a) Show that \equiv is an equivalence on G .
 (b) Describe the equivalence classes (called *double cosets*).
36. If φ is the Euler function, show that $n = \sum_{d|n} \varphi(d)$, where the sum is taken over all positive divisors of n . [Hint: Theorems 8 and 9, Section 2.4.]

2.7 GROUPS OF MOTIONS AND SYMMETRIES

Group theory began with the study of subgroups of the symmetric group S_n . In this short section we discuss some of these groups, which arise from the symmetries of geometric figures. By a **figure** we mean a finite set of points called **vertices**, some pairs of which are joined by line segments. A **motion** of a geometric figure is a permutation of its vertices that can be realized by a rigid motion in space.

Given two motions σ and τ of a figure, the composite $\sigma\tau$ is also a motion obtained by first doing τ and then σ . Similarly, σ^{-1} is a motion achieved by reversing

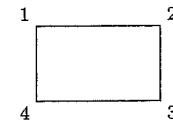
the motion that led to σ . Finally, the identity permutation ε is a motion (resulting from doing nothing at all). Hence the subgroup test gives Theorem 1.

Theorem 1. *The set of motions of a figure with n vertices is a subgroup of S_n .*

This theorem leads to many interesting groups.

Example 1. Find the group of motions of a (nonsquare) rectangle.

Solution. Label the vertices as shown. Then the motions $(1\ 2)(3\ 4)$ and $(1\ 4)(2\ 3)$ result from rotating the rectangle π radians (180°) about the vertical and horizontal axes of symmetry, respectively.



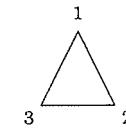
The composite of these is $(1\ 3)(2\ 4)$, which is the motion obtained by a rotation of 180° in the plane of the rectangle. Hence

$$G = \{\varepsilon, (1\ 2)(3\ 4), (1\ 4)(2\ 3), (1\ 3)(2\ 4)\}$$

is the group of motions. This group is isomorphic to the Klein group. \square

Example 2. Find the group of motions of an equilateral triangle.

Solution. Label the vertices as shown. The motions $\sigma = (1\ 2\ 3)$ and $\sigma^2 = (1\ 3\ 2)$ are achieved by clockwise rotations of $2\pi/3$ radians (120°) and $4\pi/3$ radians (240°), respectively. In addition, $\tau = (1\ 2)$ is realized by rotating the triangle π radians (180°) about the line through vertex 3 and the midpoint of the opposite side. Similarly $(1\ 3)$ and $(2\ 3)$ are motions, so the group is

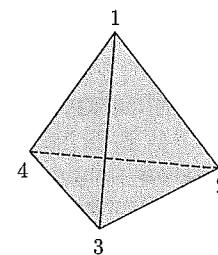


$$S_3 = \{\varepsilon, (1\ 2\ 3), (1\ 3\ 2), (1\ 2), (1\ 3), (2\ 3)\}.$$

This shows that the equilateral triangle is highly symmetric because *every* possible permutation of the vertices can be obtained by a rigid motion in space. \square

It is vital that the rigid motions allowed in Example 2 are rigid motions in *space*. If the only motions allowed were those in the plane of the triangle, the group of motions would be $\{\varepsilon, (1\ 2\ 3), (1\ 3\ 2)\}$. The permutations $(1\ 2)$, $(1\ 3)$, and $(2\ 3)$ cannot be achieved by rigid motions in the plane of the triangle.

A striking illustration of this phenomenon results when we consider the group G of motions of a tetrahedron. This figure is three-dimensional, with four vertices and six edges of equal length as in the diagram. Clearly $(1\ 2\ 3)$ is a motion of the tetrahedron, obtained by a rotation of $2\pi/3$ radians (120°) about a line through vertex 4 and the center of the opposite face. Similarly, all 3 cycles are in G . The three permutations $(1\ 2)(3\ 4)$, $(1\ 3)(2\ 4)$, and $(1\ 4)(2\ 3)$ are also motions, so $A_4 \subseteq G$, where A_4 is the alternating group of all even permutations:



$$A_4 = \left\{ \begin{array}{l} (1\ 2)(3\ 4) \\ \varepsilon \quad (1\ 4)(2\ 3) \quad (1\ 2\ 3) \quad (1\ 2\ 4) \quad (1\ 3\ 4) \quad (2\ 3\ 4) \\ (1\ 3)(2\ 4) \end{array} \right\}.$$

We claim that $A_4 = G$. Suppose on the contrary that $\sigma \in G$ is an odd motion. If $\gamma = (1 \ 2)$, write $\tau = \gamma\sigma$. Then τ is even so $\tau \in G$ and hence $\gamma = \tau\sigma^{-1}$ is in G because G is a group. But the transposition $\gamma = (1 \ 2)$ is *not* a motion of the tetrahedron, because interchanging vertices 1 and 2 by a rigid motion necessarily interchanges 3 and 4. It follows that

Example 3. The group of motions of the tetrahedron is A_4 .

This situation is analogous to that for the equilateral triangle in Example 2, where the group of motions is $A_3 = \{\varepsilon, (1 \ 2 \ 3), (1 \ 3 \ 2)\}$ if the motions are restricted to the plane containing the triangle. Any odd permutation is achieved as a motion only if the triangle is pulled out of its plane, flipped over, and placed back in its plane. Similarly, no odd permutation of the vertices of a tetrahedron can be realized by a motion in 3-space. It can be achieved only if the figure is “moved” into 4-space in the process.

Even so, these odd permutations of the vertices of a tetrahedron are *symmetries* of the figure in the intuitive sense of the word. To make this precise, we let $d(x, y)$ denote the distance between two points x and y in space. As in Section 1.4, if $\sigma \in S_n$, we write $\sigma(k) = \sigma k$ for all integers k . Given a geometric figure with n vertices labeled $1, 2, \dots, n$, a **symmetry** of the figure is a permutation σ of the vertices that preserves the distance between any two vertices; that is

$$d(\sigma k, \sigma m) = d(k, m), \quad \text{for all } k, m = 1, 2, \dots, n.$$

Clearly, any motion of a figure is a symmetry, but the converse is not true. For example, the transposition $\gamma = (1 \ 2)$ is a symmetry of the tetrahedron, but it is not a motion, as we have demonstrated.

Theorem 2. The symmetries of a figure with n vertices are a subgroup of S_n .

Proof. The identity permutation is clearly a symmetry. If σ and τ are symmetries, then for vertices k and m we have

$$d[(\sigma\tau)k, (\sigma\tau)m] = d[\sigma(\tau k), \sigma(\tau m)] = d(\tau k, \tau m) = d(k, m).$$

Hence $\sigma\tau$ is a symmetry. Finally, write $\sigma^{-1}k = k_1$ and $\sigma^{-1}m = m_1$. Then $k = \sigma k_1$ and $m = \sigma m_1$ so, since σ is a symmetry,

$$d(\sigma^{-1}k, \sigma^{-1}m) = d(k_1, m_1) = d(\sigma k_1, \sigma m_1) = d(k, m).$$

This shows that σ^{-1} is a symmetry and so completes the proof. ■

Now let G denote the group of symmetries of the tetrahedron. Then Example 3 gives $A_4 \subseteq G \subseteq S_4$, and $A_4 \neq G$ because $(1 \ 2) \in G$. This implies that $G = S_4$ because $|S_4 : A_4| = 2$ is a prime (see Example 6 §2.6). Hence

Example 4. The group of symmetries of the tetrahedron is S_4 .

The group of motions (in 3-space) of a geometric figure is thus a subgroup of the group of symmetries, and the two may be distinct as the tetrahedron shows. However, if the figure can be drawn in a plane, the two groups coincide. The reason comes from a theorem of plane geometry. Call a mapping σ from the plane to itself an **isometry** if it preserves distance; that is if $d[\sigma(x), \sigma(y)] = d(x, y)$ for all x and y . It can be shown that every isometry of the plane is a composite of translations, rotations about a point, and reflections in a line. Translations and rotations result

from motions in the plane itself, whereas reflections can only be achieved by motions in 3-space. Thus, every isometry of the plane (and hence every symmetry of a plane figure) is a motion in 3-space. Of course, this condition breaks down for a three-dimensional figure because reflections in a plane are isometries of 3-space that are not motions of 3-space.

We conclude this section by representing the dihedral group D_n as a group of motions. If $n \geq 3$, a **regular n -gon** is a plane figure with n vertices evenly placed on a circle. Thus, a regular 3-gon is an equilateral triangle, a regular 4-gon is a square, and so on. Consider the group G of all motions of a regular n -gon. There are two obvious motions:

- (1) $\sigma = (1 \ 2 \ 3 \ \cdots \ n)$ —the clockwise rotation of $2\pi/n$ radians ($360/n^\circ$) about the center of the figure.
- (2) $\tau = (1 \ n - 1)(2 \ n - 2)(3 \ n - 3) \cdots$ —the rotation of π radians (180°) about a line through the vertex n and the center of the figure.

If n is odd, then τ fixes only the vertex n , whereas if $n = 2m$, then τ fixes n and m (see Figure 2.1). If λ is any motion of the n -gon, λ is determined by its effect λ_1 and λ_2 on vertices 1 and 2. If λ_2 follows λ_1 (clockwise round the n -gon) then $\lambda = \sigma^k$ for some k . On the other hand, if λ_2 precedes λ_1 , then $\lambda = \tau\sigma^k$ for some k . For example, if $n = 7$ and the effect of λ is that shown in Figure 2.2, then λ can be achieved by $\tau\sigma^4$ as shown in Figure 2.3.

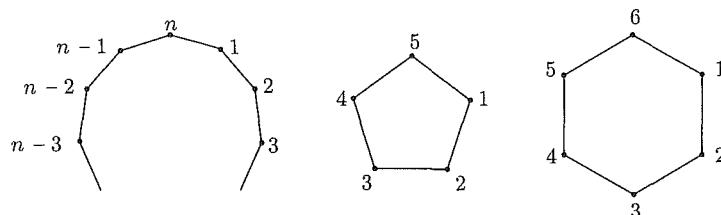


FIGURE 2.1 The difference when n is odd or even.

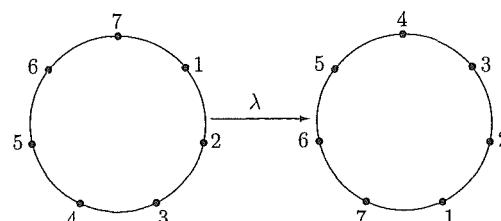
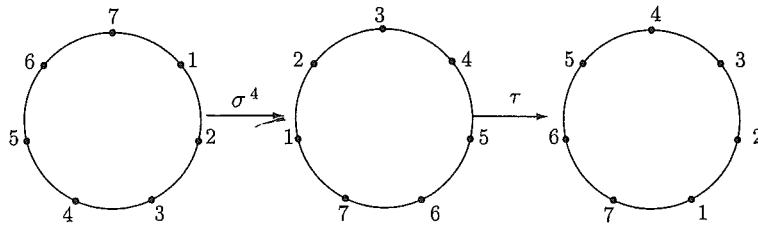


FIGURE 2.2 The effect of the motion λ .

FIGURE 2.3 The effect of $\tau\sigma^4$ is that of λ .

Because $|\sigma| = n$ and $|\tau| = 2$, it follows that

$$G = \{\varepsilon, \sigma, \sigma^2, \dots, \sigma^{n-1}, \tau, \tau\sigma, \tau\sigma^2, \dots, \tau\sigma^{n-1}\}.$$

Thus $G = \langle \sigma \rangle \cup \tau \langle \sigma \rangle$, so $|G| = 2n$. Moreover, the relation $\sigma\tau\sigma = \tau$ is valid as the following diagram shows.

$$\begin{array}{ccccccc} & \sigma & & \tau & & \sigma & \\ 1 & \rightarrow & 2 & \rightarrow & n-2 & \rightarrow & n-1 \\ 2 & \rightarrow & 3 & \rightarrow & n-3 & \rightarrow & n-2 \\ \vdots & & \vdots & & \vdots & & \vdots \\ n-1 & \rightarrow & n & \rightarrow & n & \rightarrow & 1 \\ n & \rightarrow & 1 & \rightarrow & n-1 & \rightarrow & n \end{array}$$

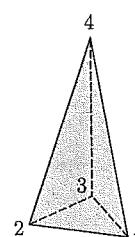
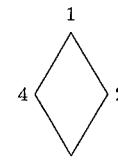
Because $|\sigma| = n$, $|\tau| = 2$, and $\sigma\tau\sigma = \tau$, the definition of D_n (in § 2.6) proves

Theorem 3. *The group of motions of a regular n -gon is isomorphic to D_n .*

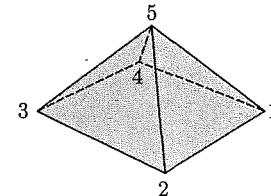
If $n = 3$, Theorem 3 shows that the group of motions of an equilateral triangle is isomorphic to D_3 , as is clear from Example 2. If $n = 4$, it shows that the group of motions of the square is isomorphic to the octic group D_4 .

Exercises 2.7

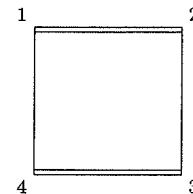
- Find the group of motions of the diamond shown—all edges, and the horizontal diagonal, of length 1.
- Describe a symmetry of the cube that is not a (three-dimensional) motion.
- Consider the figure where the base edges are of length 1 and the sloped edges are of length 2.
 - Find the group of (three-dimensional) motions.
 - Find the group of symmetries.



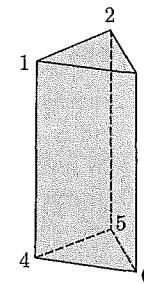
4. Consider the figure where all the edges have length 1 and the base is square.
- Find the group of (three-dimensional) motions.
 - Find the group of symmetries.



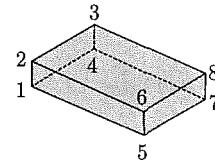
5. If the double-marked edges of the square shown are painted blue, find the subgroup of the symmetries that carry blue edges to blue edges.



6. (a) Find the group of (three-dimensional) motions of the figure where the triangle edges are of length 1 and the sides are 1×2 rectangles.
 (b) Find the group of symmetries of the figure.



7. Find the groups of motions and symmetries of the figure where each face is a nonsquare rectangle.



2.8 NORMAL SUBGROUPS

If H is a subgroup of a group G , we have seen that $aH = Ha$ may fail to hold for some $a \in G$ (Example 2 below). A subgroup H of a group G is called a **normal subgroup** of G if $gH = Hg$ holds for all g in G .

In this case H is said to be **normal** in G , written $H \triangleleft G$. These subgroups are of fundamental importance in group theory, and in this section we begin to see why.

Example 1. If G is any group, $\{1\} \triangleleft G$ because $g\{1\} = \{g\} = \{1\}g$, and $G \triangleleft G$ because $gG = G = Gg$ for all $g \in G$.

Example 2. Let $S_3 = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$, where $o(\sigma) = 3$, $o(\tau) = 2$, and $\sigma\tau\sigma = \tau$. If $H = \{\varepsilon, \sigma, \sigma^2\}$ and $K = \{\varepsilon, \tau\}$, show that $H \triangleleft S_3$ but that $K \not\triangleleft S_3$.

Solution. Clearly, $\alpha H = H = H\alpha$ for all $\alpha \in H$. Because $\sigma\tau = \tau\sigma^2$ and $\sigma^2\tau = \tau\sigma$, we get

$$H\tau = \{\tau, \sigma\tau, \sigma^2\tau\} = \{\tau, \tau\sigma^2, \tau\sigma\} = \tau H.$$

Similarly, $H\tau\sigma = \tau\sigma H$ and $H\tau\sigma^2 = \tau\sigma^2H$, so $H \triangleleft S_3$.

However, $\sigma K = \{\sigma, \sigma\tau\}$ and $K\sigma = \{\sigma, \sigma^2\tau\}$, so $\sigma K \neq K\sigma$. Hence $K \not\triangleleft S_3$. \square

Let H be a subgroup of a group G . If $g \in G$ satisfies $gh = hg$ for all $h \in H$, then obviously $gH = Hg$. In particular, this condition holds if each element h of H is in the center $Z(G)$ of G . This proves Theorems 1 and 2.

Theorem 1. *If G is a group, every subgroup of the center $Z(G)$ is normal in G . In particular, $Z(G) \triangleleft G$.*

Theorem 2. *If G is an abelian group, every subgroup of G is normal in G .*

Note that, given $g \in G$, it is *not* necessary that $gh = hg$ for all $h \in H$ to ensure that $gH = Hg$. For example, to show that $gH \subseteq Hg$ it is only necessary to show that, given $h \in H$, $gh = h'g$ for *some* $h' \in H$.

The converse of Theorem 1 is false: The subgroup H in Example 2 is normal in S_3 , but H is certainly not central in S_3 (in fact, $Z(S_3) = \{\varepsilon\}$). The converse of Theorem 2 is also false: Example 9 below exhibits a nonabelian group in which every subgroup is normal.

Example 3. If $K = \{\varepsilon, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$, show that $K \triangleleft A_4$.

Solution. If $\sigma \in S_n$ and $\gamma = (k_1 \ k_2 \ \dots \ k_r)$ is a cycle of length r , then $\sigma\gamma\sigma^{-1}$ is also a cycle of length r , in fact $\sigma\gamma\sigma^{-1} = (\sigma k_1 \ \sigma k_2 \ \dots \ \sigma k_r)$ —see Lemma 3 below. With this, let $(a \ b)(c \ d) \in K$. If $\sigma \in S_4$, then

$$\sigma[(a \ b)(c \ d)]\sigma^{-1} = \sigma(a \ b)\sigma^{-1} \ \sigma(c \ d)\sigma^{-1} = (\sigma a \ \sigma b) \ (\sigma c \ \sigma d) \in K.$$

It follows that $K \triangleleft S_4$, so certainly $K \triangleleft A_4$. \square

Theorem 3. Normality Test. *The following conditions are equivalent for a subgroup H of a group G .*

- (1) H is normal in G .
- (2) $gHg^{-1} \subseteq H$ for all $g \in G$.
- (3) $gHg^{-1} = H$ for all $g \in G$.

Proof. (1) \Rightarrow (2). Let $x \in gHg^{-1}$, say $x = ghg^{-1}$. Then $gh \in gH = Hg$ by (1), say $gh = h_1g$. Then $x = ghg^{-1} = h_1gg^{-1} = h_1 \in H$. This proves (2).

(2) \Rightarrow (3). If $g \in G$ then $gHg^{-1} \subseteq H$ by (2). Taking g^{-1} in place of g in (2), we obtain $g^{-1}Hg \subseteq H$. This implies $H \subseteq gHg^{-1}$ (verify), so $H = gHg^{-1}$, proving (3).

(3) \Rightarrow (1). Given $g \in G$, we have $gHg^{-1} = H$ by (3). If $x \in gH$, this shows that $xg^{-1} \in H$, so $x \in Hg$. This proves $gH \subseteq Hg$. Since $g^{-1}Hg = H$ by (3) (with g^{-1} replacing g), a similar argument shows that $Hg \subseteq gH$. Now (1) follows. \blacksquare

Conditions (2) and (3) in Theorem 3 become even more useful if G has a known set X of generators (see Theorem 10 §2.4). The proof is Exercise 13.

Corollary 1. *If $G = \langle X \rangle$, a subgroup H is normal in G if and only if $xHx^{-1} \subseteq H$ for all $x \in X$. In particular, $\langle a \rangle \triangleleft G$ if and only if $gag^{-1} \in \langle a \rangle$ for all $g \in G$.*

If H is a subgroup of G and $g \in G$, recall (Theorem 5 §2.3) that gHg^{-1} is also a subgroup of G which is isomorphic to H ,³¹ and is called a **conjugate** of H in G . For

³¹In fact $gHg^{-1} = \sigma_g(H)$ where $\sigma_g : G \rightarrow G$ is the inner automorphism determined by g .

this reason, normal subgroups of G are sometimes called *self-conjugate* subgroups. Incidentally, this discussion proves

Corollary 2. *If H is a subgroup of G , and if G has no other subgroups isomorphic to H , then H is normal in G .*

In particular, if H is finite and H is the *only* subgroup of its order, then $H \triangleleft G$ because $|gHg^{-1}| = |H|$ for all $g \in G$.

Theorem 3 suggests a stronger condition than normality. A subgroup H of a group G is called a **characteristic subgroup** of G if $\sigma(H) = H$ for all automorphisms $\sigma : G \rightarrow G$ (equivalently if $\sigma(H) \subseteq H$ for all automorphisms σ). The center $Z(G)$ is characteristic in G , and other examples are given in Exercise 24. If $\sigma = \sigma_a$ is the inner automorphism induced by $a \in G$ then $aHa^{-1} = \sigma_a(H)$, and it follows that characteristic subgroups are necessarily normal. However, the converse is false by Exercise 24 (c). The following result is often useful.

Corollary 3. *If $K \triangleleft G$ and $H \subseteq K$ is characteristic in K , then necessarily $H \triangleleft G$.*

Proof. If $a \in G$, $\sigma_a : K \rightarrow K$ is an automorphism of K because $K \triangleleft G$. It follows that $\sigma_a(H) = H$ because H is characteristic in K . Hence $K \triangleleft G$. ■

Corollary 3 fails if H is merely normal in K (Exercise 4). Many important subgroups are characteristic subgroups (for example, the center); some of their properties are given in Exercise 24.

Example 4. Let $G = GL_2(\mathbb{R})$ and let H be the subgroup of all matrices with determinant 1. Show that $H \triangleleft G$.

Solution. If $A \in G$ and $B \in H$, the properties of determinants give

$$\det(ABA^{-1}) = \det A \det B \det A^{-1} = \det A \cdot 1 \cdot \frac{1}{\det A} = 1.$$

This shows that $ABA^{-1} \in H$, so H is normal in G by part (2) of Theorem 3. □

Theorem 4. *If H is a subgroup of index 2 in G , then H is normal in G .*

Proof. Let $a \in G$. If $a \in H$, then $Ha = H = aH$. If $a \notin H$ then (because H has exactly 2 right cosets) $G = H \cup Ha$, a disjoint union. Hence $Ha = G \setminus H$. Similarly $aH = G \setminus H$ as H has two left cosets, so $Ha = G \setminus H = aH$. Thus, $H \triangleleft G$. ■

Note that subgroups of index 3 need not be normal (Example 2).

Example 5. Show $A_n \triangleleft S_n$ where A_n is the alternating group.

Solution. The alternating group A_n is of index two in S_n (Theorem 8 §1.4). □

Example 6. Let $D_n = \{1, a, a^2, \dots, a^{n-1}, b, ba, ba^2, \dots, ba^{n-1}\}$ denote the dihedral group, where $o(a) = n$, $o(b) = 2$, and $aba = b$. Then $\langle a \rangle \triangleleft D_n$ by Theorem 4 because $\langle a \rangle = \{1, a, \dots, a^{n-1}\}$ has index 2 in D_n . □

We defined $H \triangleleft G$ to mean $gH = Hg$ for all $g \in G$, which is a kind of commutativity condition on H . The next result gives a situation where actual commuting of elements is implied. It will be referred to later.

Lemma 1. *Let $H \triangleleft G$ and $K \triangleleft G$. If $H \cap K = \{1\}$, then $hk = kh$ for all elements $h \in H$ and $k \in K$.*

Proof. Consider $x = hkh^{-1}k^{-1}$. Thinking of $x = h(kh^{-1}k^{-1})$ we see that $x \in H$ because $kh^{-1}k^{-1} \in kHk^{-1} = H$ since $H \triangleleft G$. Similarly, writing $x = (hkh^{-1})k^{-1}$ shows that $x \in K$ because $K \triangleleft G$. Hence $x \in H \cap K = \{1\}$ by hypothesis, so $x = 1$. But then $hkh^{-1}k^{-1} = 1$, which gives $hk = kh$. ■

Let H and K be subgroups of a group G . The intersection $H \cap K$ is the *largest* subgroup of G contained in both H and K (it contains any such subgroup), and one wonders if there is a *smallest* subgroup of G containing both H and K . Note that, while $H \cup K$ is the smallest *subset* containing H and K , it is a subgroup only if $H \subseteq K$ or $K \subseteq H$ (see Exercise 17 §2.3). A much more useful construction turns out to be the **product** HK of the subgroups defined as follows:

$$HK = \{hk \mid h \in H, k \in K\}.$$

Then HK contains both H and K , and is contained in any such subgroup, but HK need not be a subgroup (consider $H = \{\varepsilon, \tau\}$ and $K = \{\varepsilon, \tau\sigma\}$ in S_3 with the usual notation). However we do have the following result.

Lemma 2. *The following are equivalent for subgroups H and K of a group G :*

- (1) HK is a subgroup of G .
- (2) $HK = KH$.
- (3) KH is a subgroup of G .

Proof. We prove only (1) \Leftrightarrow (2); then (1) \Leftrightarrow (3) follows by interchanging H and K .

(1) \Rightarrow (2). If $kh \in KH$, then $kh = (h^{-1}k^{-1})^{-1} \in HK$ by (1). This shows that $KH \subseteq HK$. On the other hand, if $hk \in HK$ then $k^{-1}h^{-1} = (hk)^{-1} \in HK$ by (1), say $k^{-1}h^{-1} = h_1k_1$. Hence $hk = k_1^{-1}h_1^{-1} \in KH$, so $HK \subseteq KH$.

(2) \Rightarrow (1). We use the subgroup test. Clearly $1 = 1 \cdot 1 \in HK$ always holds. If $hk \in HK$ then $(hk)^{-1} = k^{-1}h^{-1} \in KH = HK$ by (2). Finally, given hk and h_1k_1 in HK , we have $kh_1 \in KH = HK$, say $kh_1 = h_2k_2$. But then it follows that $(hk)(h_1k_1) = h(h_2k_2)k_1 = (hh_2)(k_2k_1) \in HK$, which completes the proof of (1). ■

A note of caution is needed here: To say that $HK = KH$ does *not* mean that $hk = kh$ for all $h \in H$ and $k \in K$. To show that $HK \subseteq KH$ means that, if $h \in H$ and $k \in K$ are given, then $hk = k_1h_1$ for *some* $h_1 \in H$ and $k_1 \in K$.

If G is abelian then HK is certainly a subgroup by Lemma 2. There is more:

Theorem 5. *Let H and K be subgroups of a group G .*

- (1) *If H or K is normal in G , then $HK = KH$ is a subgroup of G .*
- (2) *If both H and K are normal in G , then HK is also normal in G .*

Proof. (1) Suppose that K is normal in G . If $hk \in HK$ then $hk = (hkh^{-1})h \in KH$ because $hkh^{-1} \in hKh^{-1} = K$. Hence $HK \subseteq KH$. The other inclusion is proved the same way, so Lemma 2 applies. A similar argument works if $H \triangleleft G$.

(2) If $g \in G$ and $hk \in HK$, then $g^{-1}(hk)g = (g^{-1}hg)(g^{-1}kg) \in HK$ because $H \triangleleft G$ and $K \triangleleft G$. This proves (2). ■

Many groups arise as direct products of groups of smaller order, and the following useful theorem gives an important way to recognize when this is the case.

Theorem 6. If $H \triangleleft G$ and $K \triangleleft G$ satisfy $H \cap K = \{1\}$, then $HK \cong H \times K$.

Proof. First, HK is a subgroup of G by Theorem 5. Define

$$\sigma : H \times K \rightarrow HK \quad \text{by} \quad \sigma(h, k) = hk \text{ for all } h \in H \text{ and } k \in K.$$

We show that σ is an isomorphism. Given (h, k) and (h_1, k_1) in $H \times K$, we have $h_1k = kh_1$ by Lemma 1, so σ is a homomorphism because

$$\sigma[(h, k) \cdot (h_1, k_1)] = \sigma(hh_1, kk_1) = hh_1kk_1 = hk_1k_1 = \sigma(h, k) \cdot \sigma(h_1, k_1).$$

Since σ is clearly onto, it remains to show that σ is one-to-one. If $\sigma(h, k) = \sigma(h_1, k_1)$, then $hk = h_1k_1$ so $h_1^{-1}h = k_1k^{-1} \in H \cap K = \{1\}$. Thus $h_1^{-1}h = 1 = k_1k^{-1}$, so $h = h_1$ and $k = k_1$. But then $(h, k) = (h_1, k_1)$, proving that σ is one-to-one. ■

The map σ in Theorem 6 is a bijection for *any* subgroups H and K . Hence

Corollary 1. If G is a finite group, and H and K are subgroups with $H \cap K = \{1\}$, then $|HK| = |H||K|$.

Corollary 2. Let G be a finite group and let H and K be normal subgroups such that $H \cap K = \{1\}$ and $|H||K| = |G|$. Then $G \cong HK$.

Proof. By Corollary 1, we have $|HK| = |H||K| = |G|$. Hence $G = HK$ because $HK \subseteq G$ and G is finite, so Theorem 6 applies. ■

Examples 7 and 8 below illustrate how to use Theorem 6. It is easy to verify that the direct product of two cyclic groups of relatively prime orders is again cyclic (Exercise 25 §2.4). Example 7 is the converse. Recall that C_n denotes the generic cyclic group of order n .

Example 7. Let m and n be relatively prime positive integers. If G is a cyclic group of order mn , show that $G \cong C_m \times C_n$.

Solution. Let $G = \langle a \rangle$ where $o(a) = mn$, and write $H = \langle a^n \rangle$ and $K = \langle a^m \rangle$. Then $|H| = o(a^n) = m$ and $|K| = o(a^m) = n$, so $H \cong C_m$ and $K \cong C_n$. Moreover, $H \cap K = \{1\}$ by Lagrange's theorem (Corollary 4). Also, $H \triangleleft G$ and $K \triangleleft G$ because G is abelian, and $|H||K| = mn = |G|$. Hence $G \cong H \times K$ by Corollary 2 of Theorem 6, that is $G \cong C_m \times C_n$ by Example 13 §2.5. □

The fundamental theorem of finite abelian groups asserts that every finite abelian group is isomorphic to a uniquely determined direct product of cyclic groups. We prove this assertion in Section 7.2. Example 8 gives a special case.

Example 8. Let G be an abelian group and assume that $|G| = p^2$, p a prime. Then either $G \cong C_{p^2}$ is cyclic or $G \cong C_p \times C_p$.

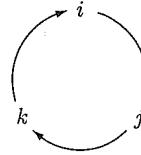
Solution. Assume that G is not cyclic. Then $o(g) = p$ if $1 \neq g \in G$ because $o(g)$ divides p^2 . Choose $a \in G$ with $o(a) = p$, and write $H = \langle a \rangle \cong C_p$. Then $H \neq G$, so choose $b \notin H$, $b \in G$, and write $K = \langle b \rangle \cong C_p$. Hence $|K| = p$ too, so we have $|H||K| = p \cdot p = |G|$. Moreover $H \triangleleft G$ and $K \triangleleft G$ because G is abelian. Finally, $H \cap K = \{1\}$ because, otherwise, $H = H \cap K = K$ by Corollary 3 of Lagrange's theorem. Thus $G \cong H \times K$ by Corollary 2 of Theorem 6, that is $G \cong C_p \times C_p$. □

We have already noted (Theorem 2) that every subgroup of an abelian group is normal. The converse is not true: A nonabelian group of order 8 exists in which every subgroup is normal. It is constructed as follows: Let

$$Q = \{\pm 1, \pm i, \pm j, \pm k\}$$

be a set of eight elements with multiplication determined by the following equations:

$$\begin{aligned} i^2 &= j^2 = k^2 = ijk = -1, \\ ij &= k = -ji, \\ jk &= i = -kj, \\ ki &= j = -ik. \end{aligned}$$



Here 1 and -1 multiply as usual, and the multiplication of i, j , and k is best remembered by the diagram above: The product of any two of i, j , and k taken clockwise around the circle is the next one, whereas the product counterclockwise is the negative of the next one. One realization of Q is in $GL_2(\mathbb{C})$ where, if $w \in \mathbb{C}$ satisfies $w^2 = -1$, we take

$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, i = \begin{bmatrix} w & 0 \\ 0 & -w \end{bmatrix}, j = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \text{ and } k = \begin{bmatrix} 0 & w \\ w & 0 \end{bmatrix}.$$

The group Q in the preceding discussion is called the **quaternion group**, and it rules out the converse to Theorem 2.

Example 9. Show that Q is a nonabelian group in which every subgroup is normal.

Solution. Q is nonabelian because $ij = k$ while $ji = -k$. If a subgroup H contains one of $\pm i$, $\pm j$, or $\pm k$, then $|H| = 4$ or 8 (because these elements have order 4), so $H \triangleleft Q$ by Theorem 4. Otherwise, $H \subseteq \{1, -1\} \subseteq Z(Q)$ and again $H \triangleleft Q$. \square

Simple Groups

Lagrange's theorem shows that the cyclic groups G of prime order have no subgroups except $\{1\}$ and G . More generally, if G is a group then we say that

G is **simple** if $G \neq \{1\}$ and the only *normal* subgroups of G are $\{1\}$ and G .

Theorem 7. An abelian group $G \neq \{1\}$ is simple if and only if it is cyclic of prime order.

Proof. Suppose G is simple and abelian. Then the only subgroups of G are $\{1\}$ and G (all subgroups are normal). If $a \in G$, $a \neq 1$, this means that $\langle a \rangle = G$. Then $o(a) \neq \infty$ because, otherwise, $\langle a^2 \rangle$ does not equal $\{1\}$ or G . So $o(a)$ is finite, say $o(a) = n \geq 2$. If $p|n$ for some prime p , then $\langle a^{n/p} \rangle$ is a subgroup of order p by Theorem 5 §2.4. Hence $G = \langle a^{n/p} \rangle$ is cyclic of prime order.

The converse is by Lagrange's theorem. \blacksquare

Nonabelian finite simple groups are more difficult to find. We conclude with a proof that, although A_4 is not simple by Example 3, the alternating groups A_n are all simple if $n \geq 5$. This has applications in the theory of equations (Chapter 10). The proof requires three preliminary results, the first two of independent interest.

Lemma 3. If $\sigma \in S_n$ and $\gamma = (k_1 \ k_2 \ \cdots \ k_r)$ is a cycle of length r , then $\sigma\gamma\sigma^{-1}$ is also a cycle of length r . In fact, $\sigma\gamma\sigma^{-1} = (\sigma k_1 \ \sigma k_2 \ \cdots \ \sigma k_r)$.

Proof. Write $\delta = (\sigma k_1 \ \sigma k_2 \ \cdots \ \sigma k_r)$. Because σ is one-to-one, δ is a cycle of length r . We must show that $\sigma\gamma\sigma^{-1} = \delta$, that is $\sigma\gamma = \delta\sigma$, that is $\sigma(\gamma k) = \delta(\sigma k)$ for each $k = 1, 2, \dots, n$. The reader can verify that (writing $k_{r+1} = k_1$)

$$\sigma(\gamma k_i) = \sigma k_{i+1} = \delta(\sigma k_i) \text{ for each } i = 1, 2, \dots, r.$$

But $\sigma(\gamma k) = \sigma k = \delta(\sigma k)$ whenever $k \notin \{k_1, k_2, \dots, k_r\}$ because (as σ is one-to-one) this implies that $\sigma k \notin \{\sigma k_1, \sigma k_2, \dots, \sigma k_r\}$. This is what we wanted. ■

Lemma 4. If $n \geq 2$, A_n is generated by the 3-cycles.

Proof. As each 3-cycle is in A_n , it suffices to show that each permutation $\sigma \in A_n$ is either ε or a product of 3-cycles. But σ is even and so is a product of pairs of transpositions. Hence the following formulas complete the proof:

$$(i \ j)(i \ j) = \varepsilon, \quad (i \ j)(i \ k) = (i \ k \ j), \quad \text{and} \quad (i \ j)(k \ l) = (i \ l \ k)(i \ j \ k). \quad \blacksquare$$

Lemma 5. Suppose that $n \geq 5$. If $H \triangleleft A_n$ and H contains a 3-cycle, then $H = A_n$.

Proof. If $(i \ j \ k) \in H$ we claim that $(a \ j \ k)$, $(i \ a \ k)$ and $(i \ j \ a)$ are all in H for any $a \notin \{i, j, k\}$. Indeed: Since $n \geq 5$, choose $b \notin \{a, i, j, k\}$. Then we have $(a \ j \ k) = (i \ a \ b)(i \ j \ k)(i \ a \ b)^{-1} \in H$ by Lemma 3. Similarly, $(i \ a \ k), (i \ j \ a) \in H$.

Let $\sigma = (1 \ 2 \ 3) \in H$: by Lemma 4 we must show that every 3-cycle $\tau = (i \ j \ k)$ is in H . Write $S = \{1, 2, 3\}$ and $T = \{i, j, k\}$. If $|S \cap T| = 3$, then $S = T$ so $\tau = \sigma \in H$ or $\tau = \sigma^{-1} \in H$. If $|S \cap T| = 2$, say $i = 1$, $j = 2$ and $k \neq 3$, then $(1 \ 2 \ 3) \in H \Rightarrow \tau = (1 \ 2 \ k) \in H$ by the first paragraph. If $|S \cap T| = 1$, say $i = 1$ and $\{2, 3\} \cap \{j, k\} = \emptyset$, then $(1 \ 2 \ 3) \in H \Rightarrow (1 \ j \ 3) \in H \Rightarrow \tau = (1 \ j \ k) \in H$. Finally if $|S \cap T| = 0$ then $(1 \ 2 \ 3) \in H \Rightarrow (i \ 2 \ 3) \in H \Rightarrow (i \ j \ 3) \in H \Rightarrow \tau = (i \ j \ k) \in H$. ■

Theorem 8. If $n \geq 5$, the alternating group A_n is simple.

Proof. Let $H \triangleleft A_n$, $H \neq \{\varepsilon\}$. Among all elements of H (excluding ε) let τ be one that moves the smallest number m of integers. Then $m \geq 3$, because $\tau \in A_n$ is not a transposition. If $m = 3$ then τ is a 3-cycle, and we are done by Lemma 5. So assume $m \geq 4$; we show that this leads to a contradiction. Factor τ into disjoint cycles and consider two cases.

- *Case 1: τ contains a cycle of length ≥ 3 , say $\tau = (1 \ 2 \ 3 \ \cdots) \gamma_2 \cdots \gamma_r$.* If τ moves exactly 4 integers, then $\tau = (1 \ 2 \ 3 \ k)$ is odd. So assume that τ moves (say) 4 and 5, as well as 1, 2, and 3. Let $\beta = (3 \ 4 \ 5)$ and write $\tau_1 = \tau^{-1} \beta \tau \beta^{-1}$. Then $\tau_1 \in H$ because $H \triangleleft A_n$, and $\tau_1 \neq \varepsilon$ because $\tau_1 2 = \tau^{-1} 4 \neq 2$. Moreover, if $k > 5$ is fixed by τ , then k is also fixed by τ_1 (because $\beta k = k$). Hence if τ_1 moves $k > 5$, then τ also moves k . But τ_1 fixes 1, whereas τ does not. Thus τ_1 moves fewer elements than τ , a contradiction.
- *Case 2: τ is a product of disjoint transpositions, say, $\tau = (1 \ 2)(3 \ 4) \cdots$.* As before, let $\beta = (3 \ 4 \ 5)$ and $\tau_1 = \tau^{-1} \beta \tau \beta^{-1}$. Now τ_1 fixes 1 and 2 and any integer $k > 5$ that is fixed by τ . Because $\tau_1 \neq \varepsilon$ (for example, $\tau_1 5 = 3$), this is a contradiction, as in Case 1. ■

Other infinite families of finite simple groups exist (in addition to the alternating groups A_n , $n \geq 5$). The complete classification of these groups was first given in 1981. This was the culmination of more than 30 years of effort by hundreds of mathematicians, yielding thousands of pages of published work. It is certainly one of the greatest achievements of twentieth-century mathematics. One spectacular

landmark came in 1963 when J. Thompson and W. Feit proved³² a long-standing conjecture of William Burnside that every finite nonabelian simple group has even order (the proof is more than 250 pages long!). Thompson went on to publish the “N-group” paper in which he introduced many fundamental techniques, and which has been called the single most important paper in simple group theory.³³ Then in the 1970s, M. Aschbacher carried the work forward in a series of papers, building on the methods of Thompson. The main difficulty was the existence of *sporadic* finite simple groups not belonging to any of the known families. R. L. Griess finally constructed the largest of these, called the *monster* (the order is approximately 8×10^{53}). The complete classification encompasses several infinite families of finite simple groups and exactly 26 sporadic groups.³⁴

Exercises 2.8

1. Consider $D_{12} = \{1, a, \dots, a^{11}, b, ba, \dots, ba^{11}\}$, where $o(a) = 12$, $o(b) = 2$, $aba = b$. In each case show that H is a subgroup of D_{12} and determine if $H \triangleleft D_{12}$.
 - (a) $H = \{1, a^6, b, ba^6\}$
 - (b) $H = \{1, a^4, a^8, b, ba^4, ba^8\}$
 - (c) $H = \{1, a^2, a^4, a^6, a^8, a^{10}, b, ba^2, ba^4, ba^6, ba^8, ba^{10}\}$
2. Find all normal subgroups of D_4 . [Hint: Exercise 7 below.]
3. Let $K = \{\varepsilon, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$. Show that K is the only normal subgroup of A_4 apart from A_4 and $\{\varepsilon\}$. [Hint: Exercise 34 §2.6.]
4. If $D_4 = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$, $K = \{1, b\}$ and $H = \{1, a^2, b, ba^2\}$ show that $K \triangleleft H \triangleleft D_4$, but $K \not\triangleleftharpoonup D_4$.
5. If $K \triangleleft H$ and $H \triangleleft G$, show that $aKa^{-1} \triangleleft H$ for all $a \in G$. (See Theorem 5 §2.3.)
6. Let H be a subgroup of a group G . If for each $a \in G$ there exists $b \in G$ such that $aH = Hb$, show that $H \triangleleft G$.
7. If $H \triangleleft G$ and $|H| = 2$, show that $H \subseteq Z(G)$. Is this true when $|H| = 3$?
8. If H is a subgroup of G and $K \triangleleft G$, show that $H \cap K \triangleleft H$. Is $H \cap K \triangleleft K$?
9. Given a group G , let $D = \{(g, g) \mid g \in G\}$. Show that D is a normal subgroup of $G \times G$ if and only if G is abelian.
10. Let $N \triangleleft G$ and $K \triangleleft G$. Show that $N \cap K \triangleleft G$.
11. Let p and q be distinct primes. If G is a group of order pq that has a unique subgroup of order p and a unique subgroup of order q , show that G is cyclic. [Hint: Corollary 2 of Theorem 6 and Exercise 25 §2.4.]
12. Let $K \triangleleft G$ where K is cyclic. Show that every subgroup of K is normal in G .
13. Let X be a nonempty subset of a group G .
 - (a) If $G = \langle X \rangle$ (see Theorem 10 § 2.4) and H is a subgroup of G , show that $H \triangleleft G$ if and only if $x^{-1}Hx \subseteq H$ for all $x \in X$.
 - (b) Show that $\langle X \rangle$ is normal in G if and only if $gXg^{-1} \subseteq \langle X \rangle$ for all $g \in G$.

³²Thompson, J. G., and Feit, W., Solvability of Groups of Odd Order, *Pacific Journal of Mathematics*, 13 (1963), 775–1029.

³³John Thompson was awarded the Fields Medal in 1970, the highest honor a mathematician can receive.

³⁴More information can be found in Chapter 17 of “Finite Groups” by D. Gorenstein, Chelsea, 1980.

14. If $G = H \times K$ is finite, find $H_1 \triangleleft G$ and $K_1 \triangleleft G$ such that $H_1 \cong H$, $K_1 \cong K$, $H_1 \cap K_1 = \{1\}$, and $|G| = |H_1| \cdot |K_1|$. (Converse of Theorem 6.)
15. Let K be a subgroup of G of index 2.
- If $a \in G \setminus K$ and $b \in G \setminus K$, show that $ab \in K$.
 - If H is a subgroup of G and $H \not\subseteq K$, show that $|H : H \cap K| = 2$. [Hint: If $h_0 \in H \setminus K$, show that $h \mapsto hh_0$ is a bijection $H \cap K \rightarrow H \setminus (H \cap K)$.]
16. Show that $\text{inn } G \triangleleft \text{aut } G$ for any group G .
17. Let $D_n = \{1, a, \dots, a^{n-1}, b, ba, \dots, ba^{n-1}\}$ with $o(a) = n$, $o(b) = 2$, and $aba = b$.
- Show that every subgroup K of $\langle a \rangle$ is normal in D_n .
 - If n is odd and $K \triangleleft D_n$, show that $K = D_n$ or $K \subseteq \langle a \rangle$.
18. (a) Let Q denote the quaternion group. If $a = i$ and $b = j$ show that Q has the form $Q = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$ where $o(a) = 4$, $aba = b$, and $b^2 = a^2$. Show further that these conditions determine the Cayley table of Q .
- (b) If G is a nonabelian group of order 8, show that $G \cong D_4$ or $G \cong Q$. [Hint: See Theorem 3 §2.6; use Theorem 4 and (a).]
19. If H and K are subgroups of G , show that HK is a subgroup if and only if $HK \subseteq KH$, if and only if $KH \subseteq HK$.
20. If $G = HK$ where H and K are subgroups such that $hk = kh$ for all $h \in H$ and $k \in K$, show that $H \triangleleft G$ and $K \triangleleft G$.
21. If $H \subseteq G$ and $K \subseteq G$ are subgroups with $HK = KH$, show that $HK = \langle H \cup K \rangle$. (See Theorem 8 §2.4.)
22. Let G be a group with $|G| = mn$ where m and n are relatively prime, and let H and K be subgroups where $|H| = m$, $|K| = n$. If $hk = kh$ for all $h \in H$ and $k \in K$, show that $G \cong H \times K$.
23. (a) Let $n = 2m$, where m is odd. Show that $D_n \cong C_2 \times D_m$, where C_2 is cyclic of order 2. [Hint: Corollary 2 of Theorem 6.]
- (b) Is $D_{12} \cong C_3 \times D_4$? Justify your answer.
24. A subgroup H of a group G is called a **characteristic** subgroup if $\sigma(H) \subseteq H$ for all automorphisms σ of G .
- Show that every characteristic subgroup is normal.
 - Show that if H is characteristic in G then $\sigma(H) = H$ for all $\sigma \in \text{aut } G$.
 - If $G = C_2 \times C_2$ show that $H = C_2 \times \{1\}$ is normal in G but not characteristic. [Hint: Consider $\sigma : G \rightarrow G$ given by $\sigma(x, y) = (y, x)$.]
 - Show that the center $Z(G)$ is characteristic in G .
 - If $H \subseteq K \triangleleft G$ and H is characteristic in K , show that $H \triangleleft G$.
 - If K is characteristic in H and H is characteristic in G , show that K is characteristic in G . (Compare with Exercise 4.)
 - Show that every subgroup of a cyclic group G is characteristic in G . Is this true if G is merely abelian?
 - If H and K are characteristic in G , show that $H \cap K$ is characteristic in G .
 - If H is a subgroup of G , let $K = \{g \in G \mid g \in \sigma(H) \text{ for all } \sigma \in \text{aut } G\}$. Show that K is characteristic in G , that $K \subseteq H$, and that K contains every characteristic subgroup of G that is contained in H .
25. If X is a nonempty subset of a group G , define the **normalizer** $N(X)$ of X by $N(X) = \{a \in G \mid aXa^{-1} = X\}$.
- Show that $N(X)$ is a subgroup of G .
 - If H is a subgroup of G , show that $H \triangleleft N(H)$.

- (c) If H is a subgroup of G , show that $N(H)$ is the largest subgroup of G in which H is normal. That is, if $H \triangleleft K$, and K is a subgroup of G , then $K \subseteq N(H)$.
26. If H is a subgroup of G , define the **core** of H , denoted $\text{core } H$, to be the intersection of all the conjugates of H in G ; that is,

$$\text{core } H = \{g \in G \mid g \in aHa^{-1} \text{ for all } a \in G\} = \bigcap \{aHa^{-1} \mid a \in G\}.$$

- (a) Show that $\text{core } H \triangleleft G$ and $\text{core } H \subseteq H$.
 (b) Show that $\text{core } H$ is the largest normal subgroup of G that is contained in H ; that is, if $K \triangleleft G$ and $K \subseteq H$, then $K \subseteq \text{core } H$.
 (c) Show that $\text{core}(H \cap K) = \text{core } H \cap \text{core } K$ for all subgroups H and K .
27. If X is a nonempty subset of a group G , define the **normal closure** \bar{X} of X to be the intersection of all normal subgroups of G that contain X ; that is,

$$\bar{X} = \{g \in G \mid g \in N \text{ for all } N \triangleleft G, X \subseteq N\} = \bigcap \{N \mid X \subseteq N \triangleleft G\}.$$

- (a) Show that $\bar{X} \triangleleft G$ and $X \subseteq \bar{X}$.
 (b) Show that \bar{X} is the smallest normal subgroup of G that contains X ; that is, $X \subseteq N$ and $N \triangleleft G$ implies that $\bar{X} \subseteq N$.
 (c) Show that $\overline{H \cap K} \subseteq \bar{H} \cap \bar{K}$ for all subgroups H and K of G , and that this need not be equality.
 28. If X is a nonempty subset of a group G , define the **centralizer** $C(X)$ of X by $C(X) = \{c \in G \mid cx = xc \text{ for all } x \in X\}$. Note that $C(G) = Z(G)$.
 (a) Show that $C(X)$ is a subgroup of G .
 (b) If $K \triangleleft G$, show that $C(K) \triangleleft G$.

2.9 FACTOR GROUPS

If $n \geq 2$, recall the construction of \mathbb{Z}_n in Section 1.3. Given the subgroup $n\mathbb{Z}$ of $(\mathbb{Z}, +)$, the set \mathbb{Z}_n consists of all “residue classes” $\bar{a} = \{x \in \mathbb{Z} \mid x \equiv a \pmod{n}\}$ where $a \in \mathbb{Z}$. These classes are really cosets $\bar{a} = n\mathbb{Z} + a$. Moreover, we defined addition in \mathbb{Z}_n by $\bar{a} + \bar{b} = \overline{a + b}$; that is,

$$(n\mathbb{Z} + a) + (n\mathbb{Z} + b) = n\mathbb{Z} + (a + b).$$

This suggests a general definition: If K is a subgroup of a multiplicative group G , we could define an analogous multiplication on the set of right cosets by

$$KaKb = Kab \quad \text{for all } a, b \in G. \tag{*}$$

However, this may not make sense for some subgroups K because cosets can have different generators: $Ka = Ka_1$ can happen where a and a_1 may not be equal.

More precisely, let $x = Ka = Ka_1$ and $y = Kb = Kb_1$ be cosets. If we multiply $x = Ka$ and $y = Kb$ using $(*)$ we get $xy = Kab$, but if we view x and y as $x = Ka_1$ and $y = Kb_1$ we obtain $xy = Ka_1b_1$. Clearly, what is needed is:

$$\text{If } Ka = Ka_1 \text{ and } Kb = Kb_1 \text{ then necessarily } Kab = Ka_1b_1.$$

In this case we say that the multiplication $KaKb = Kab$ is **well defined**. This condition on K is equivalent to K being normal in G .

Lemma. *The following conditions are equivalent for a subgroup K of G .*

- (1) K is normal in G .
- (2) $KaKb = Kab$ is a well defined multiplication of right cosets.

Proof. (1) \Rightarrow (2). If $K \triangleleft G$, let $Ka = Ka_1$ and $Kb = Kb_1$, that is $aa_1^{-1} \in K$ and $bb_1^{-1} \in K$. We must show that $Kab = Ka_1b_1$, that is $ab(a_1b_1)^{-1} \in K$. Compute

$$ab(a_1b_1)^{-1} = ab(b_1^{-1}a_1^{-1}) = a(bb_1^{-1})a_1^{-1} = [a(bb_1^{-1})a^{-1}](aa_1^{-1}) \in K$$

because $aKa^{-1} \subseteq K$. This is what we wanted.

(2) \Rightarrow (1). If $a \in G$ we must show that $aka^{-1} \in K$ for all $k \in K$. Clearly $Ka = Ka$ and $Kk = K1$, so applying (2) gives $Kak = Kal$, that is $Kak = Ka$. But then $(ak)a^{-1} \in K$, as required. ■

Theorem 1. Let $K \triangleleft G$ and write $G/K = \{Ka \mid a \in G\}$ for the set of cosets.

- (1) G/K is a group under the operation $KaKb = Kab$.
- (2) The mapping $\varphi : G \rightarrow G/K$ given by $\varphi(a) = Ka$ is an onto homomorphism.
- (3) If G is abelian, then G/K is abelian.
- (4) If $G = \langle a \rangle$ is cyclic, then G/K is also cyclic; in fact, $G/K = \langle Ka \rangle$.
- (5) If $|G : K|$ is finite then $|G/K| = |G : K|$; if $|G|$ is finite then $|G/K| = \frac{|G|}{|K|}$.

Proof. (1) The operation on G/K is well defined by the Lemma. The unity of G/K is $K = K1$ because $KaK1 = Ka = K1Ka$ for all Ka in G/K . We have $KaKa^{-1} = K1 = Ka^{-1}Ka$, so the inverse of the coset Ka is $(Ka)^{-1} = Ka^{-1}$. Finally, associativity in G/K is inherited from G :

$$Ka(KbKc) = KaKbc = Ka(bc) = K(ab)c = KabKc = (KaKb)Kc$$

(2) We have $\varphi(a)\varphi(b) = KaKb = Kab = \varphi(ab)$ for all $a, b \in G$, so φ is a homomorphism. It is clearly onto.

(3) If G is abelian, $KaKb = Kab = Kba = KbKa$, proving (3).

(4) Let $G = \langle a \rangle = \{a^k \mid k \in \mathbb{Z}\}$, so every coset in G/K has the form Ka^k for some integer k . If φ is the map in (2), then $Ka^k = \varphi(a^k) = \varphi(a)^k = (Ka)^k$ by Theorem 1 §2.5. It follows that $G/K = \langle Ka \rangle$, as required.

(5) As $|G : K|$ is finite, $|G/K| = |G : K|$ is the definition of the index $|G : K|$. If $|G|$ is finite, then $|G/K| = |G|/|K|$ is (2) of Lagrange's theorem. ■

Thus, if $n \geq 2$ in \mathbb{Z} then $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$ as additive groups. If K is a normal subgroup of a group G , write $G/K = \{Ka \mid a \in G\}$ as in Theorem 1. We make two definitions:

The group G/K of all cosets of K in G is called the **factor group** of G by K .

The homomorphism $\varphi : G \rightarrow G/K$ where $\varphi(a) = Ka$ is called the **coset map**.

Hence the unity of G/K is $K = K1$, and inverses are given by $(Ka)^{-1} = Ka^{-1}$. It is important for a student of group theory (and ring theory for that matter) to develop skill in working with factor groups. All the group theoretic techniques we have developed up to now apply to these groups; the only new aspect is that the elements are now cosets.

Example 1. If G is a group then we always have $G \triangleleft G$ and $\{1\} \triangleleft G$.

If $K = G$ there is only one coset, so $G/K = \{G\}$ is the group with one element.

If $K = \{1\}$, then $Ka = \{a\}$ for each $a \in G$, so G/K is the set of all singleton subsets of G . The operation is $\{a\}\{b\} = \{ab\}$, so $G/K \cong G$ in this case.

Example 2. Let $G = \langle a \rangle$ where $o(a) = 12$, and let $K = \langle a^4 \rangle$. Find all the cosets in G/K and write down the Cayley table.

Solution. Note first that $K \triangleleft G$ because G is abelian. The cosets are

$$K = \{1, a^4, a^8\}, \quad Ka = \{a, a^5, a^9\}, \quad Ka^2 = \{a^2, a^6, a^{10}\}, \quad Ka^3 = \{a^3, a^7, a^{11}\}.$$

Two computations are needed to fill in the Cayley table: $Ka \cdot Ka^3 = K$ (because $a^4 \in K$) and $Ka^2 \cdot Ka^3 = Ka^5 = Ka$ (because $a^5 \in Ka$). Then

G/K	K	Ka	Ka^2	Ka^3
K	K	Ka	Ka^2	Ka^3
Ka	Ka	Ka^2	Ka^3	K
Ka^2	Ka^2	Ka^3	K	Ka
Ka^3	Ka^3	K	Ka	Ka^2

We have $G/K = \langle Ka \rangle$ as $Ka^2 = (Ka)^2$, $Ka^3 = (Ka)^3$, and $K = Ka^4 = (Ka)^4$. This confirms Theorem 1(4) in this case. \square

Example 3. Let $K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$. Show that $K \triangleleft A_4$, find all the cosets in A_4/K , and write down the Cayley table.

Solution. We showed that $K \triangleleft A_4$ in Example 3 §2.8. The cosets are

$$\begin{aligned} K\varepsilon &= \varepsilon K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\} \\ K(1 \ 2 \ 3) &= (1 \ 2 \ 3)K = \{(1 \ 2 \ 3), (2 \ 4 \ 3), (1 \ 4 \ 2), (1 \ 3 \ 4)\} \\ K(1 \ 3 \ 2) &= (1 \ 3 \ 2)K = \{(1 \ 3 \ 2), (1 \ 4 \ 3), (2 \ 3 \ 4), (1 \ 2 \ 4)\} \end{aligned}$$

The Cayley table is as shown.

A_4/K	K	$K(1 \ 2 \ 3)$	$K(1 \ 3 \ 2)$
K	K	$K(1 \ 2 \ 3)$	$K(1 \ 3 \ 2)$
$K(1 \ 2 \ 3)$	$K(1 \ 2 \ 3)$	$K(1 \ 3 \ 2)$	K
$K(1 \ 3 \ 2)$	$K(1 \ 3 \ 2)$	K	$K(1 \ 2 \ 3)$

Here the fact that $K(1 \ 3 \ 2) = [K(1 \ 2 \ 3)]^2$ shows that $G/K = \langle K(1 \ 2 \ 3) \rangle$ is cyclic. Of course, this also follows from the fact that $|G/K| = 3$ is prime. \square

Example 4. Consider the octic group $D_4 = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$, where $o(a) = 4$, $o(b) = 2$, and $aba = b$. Show that $Z(D_4) = \{1, a^2\}$ and that $D_4/Z(D_4)$ is isomorphic to the Klein group K_4 .

Solution. Write $Z = Z(D_4)$ for short. We have $a(ba^k) = ba^{k+3} \neq ba^{k+1} = (ba^k)a$ for each k , so $ba^k \notin Z$. Similarly, $ab = ba^3 \neq ba$ and $a^3b = ba \neq ba^3$ show that $a \notin Z$ and $a^3 \notin Z$. Hence $Z \subseteq \{1, a^2\}$. However, $a^2b = ba^2$, so a^2 commutes with both generators a and b of D_4 . This implies that $a^2 \in Z$, so $Z = \{1, a^2\}$.

Of course, $Z = Z(D_4)$ is normal in D_4 , and the cosets are

$$Z = \{1, a^2\}, \quad Za = \{a, a^3\}, \quad Zb = \{b, ba^2\}, \quad \text{and} \quad Zba = \{ba, ba^3\}.$$

Thus $D_4/Z = \{Z, Za, Zb, Zba\}$, and the Cayley table is as shown.

D_4/Z	Z	Za	Zb	Zba
Z	Z	Za	Zb	Zba
Za	Za	Z	Zba	Zb
Zb	Zb	Zba	Z	Za
Zba	Zba	Zb	Za	Z

This is evidently the noncyclic group of order 4, that is $D_4/Z \cong K_4$. \square

Example 5. Let $G = \langle a \rangle$, where $o(a) = 18$, and let $K = \langle a^6 \rangle$. Find the order of the element Ka^5 in G/K .

Solution 1. As $K = \{1, a^6, a^{12}\}$, we have $|G/K| = |G|/|K| = 18/3 = 6$. Then, from Lagrange's theorem, the order $o(Ka^5)$ of Ka^5 is 1, 2, 3, or 6. Now

$$Ka^5 \neq K, \quad (Ka^5)^2 = Ka^{10} \neq K, \quad \text{and} \quad (Ka^5)^3 = Ka^{15} \neq K.$$

Hence $o(Ka^5)$ is not 1, 2, or 3, so it must be 6.

Solution 2. $G = \langle a^5 \rangle$ because $\gcd(5, 18) = 1$, so $G = \langle a^5 \rangle$. Hence $G/K = \langle Ka^5 \rangle$ by Theorem 1. Because $|G/K| = 6$, this means that $o(Ka^5) = 6$. \square

The next theorem provides a useful method of proving that a group is abelian. We include it for reference later.

Theorem 2. Suppose a group G has a subgroup $K \subseteq Z(G)$ such that G/K is cyclic. Then G is abelian.

Proof. Let $G/K = \langle Kg \rangle$. If $a, b \in G$, this means that we can write Ka and Kb in the form $Ka = Kg^m$ and $Kb = Kg^n$. Thus $a = kg^m$ and $b = k_1g^n$ where k and k_1 are in K (and hence are central in G by hypothesis). But then

$$ab = (kg^m)(k_1g^n) = kk_1g^{m+n} = k_1kg^{n+m} = (k_1g^n)(kg^m) = ba.$$

This shows that G is abelian. \blacksquare

The Derived Subgroup

If $H \triangleleft G$ there is a useful test for determining when a factor group G/H is abelian. To motivate it, consider the following way of deciding that two cosets Ha and Hb commute:

$$HaHb = HbHa \Leftrightarrow Hab = Hba \Leftrightarrow ab(ba)^{-1} \in H \Leftrightarrow aba^{-1}b^{-1} \in H.$$

With this in mind, an element in a group G of the form $aba^{-1}b^{-1}$ is called a **commutator** and is denoted

$$[a, b] = aba^{-1}b^{-1}, \quad \text{for any } a, b \text{ in } G.$$

Hence, if $H \triangleleft G$ then G/H is abelian if and only if H contains every commutator.

The **commutator subgroup** or **derived subgroup** G' of G is defined by

$$G' = \{\text{all finite products of commutators from } G\}.$$

To see that G' really is a subgroup of G , note that $1 = [a, a]$ is in G' and that G' is clearly closed under the operation of G . The fact that G' is closed under inverses follows from the first of the following easily verified properties of commutators.

- (1) $[a, b]^{-1} = [b, a]$.
- (2) $g[a, b]g^{-1} = [gag^{-1}, gbg^{-1}]$ for all $g \in G$.

These facts reveal the relationship between G' and the abelian factor groups of G .

Theorem 3. Let G be a group and let H be a subgroup of G .

- (1) G' is a normal subgroup of G and G/G' is abelian.
- (2) $G' \subseteq H$ if and only if H is normal in G and G/H is abelian.

Proof. We have already established that G' is a subgroup of G . Since (2) \Rightarrow (1) by taking $H = G'$, we prove only (2). If $H \triangleleft G$, the above argument shows that G/H is abelian if and only if every commutator belongs to H , that is, if and only if $G' \subseteq H$. Hence it remains to show that $G' \subseteq H$ implies that $H \triangleleft G$. If $G' \subseteq H$, let $g \in G$ and $h \in H$. Then

$$ghg^{-1} = (ghg^{-1}h^{-1})h = [g, h]h \in G'h \subseteq Hh = H.$$

Thus $gHg^{-1} \subseteq H$, so $H \triangleleft G$ as required. \blacksquare

Hence $G' = \{1\}$ if and only if G is abelian. Since G is abelian if and only if $Z(G) = G$, this contrasts the way G' and $Z(G)$ measure the commutativity of the group G .

Theorem 3 asserts that G' is the *smallest* normal subgroup H of G with the property that the factor group G/H is abelian. This fact can be very useful in computing G' , as Example 6 illustrates.

Example 6. Compute D'_4 , where $D_4 = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$, where $o(a) = 4$, $o(b) = 2$, and $aba = b$.

Solution. In Example 4 we showed that the center of D_4 is $Z = \{1, a^2\}$ and that D_4/Z is abelian. Hence $D'_4 \subseteq Z$ by Theorem 3 and so, because $|Z| = 2$, either $D'_4 = \{1\}$ or $D'_4 = Z$. But $D'_4 = \{1\}$ is impossible because $D_4/\{1\} \cong D_4$ is not abelian. Hence $D'_4 = Z = Z(D_4)$. \square

Exercises 2.9

1. In each case find the cosets in G/K , write down the Cayley table of G/K , and describe the group G/K .
 - (a) $G = D_6$ and $K = Z(D_6)$
 - (b) $G = Q$ and $K = Z(Q)$
 - (c) $G = A \times B$, A and B arbitrary groups, and $K = \{(a, 1) \mid a \in A\}$
 - (d) $G = \langle a \rangle \times \langle b \rangle$, where $o(a) = 8$ and $o(b) = 2$, and $K = \langle (a^2, b) \rangle$
2. An integer n is called an exponent for a group G if $g^n = 1$ for every g in G . If n is an exponent for G ; show that it is an exponent for every factor group G/K .
3. If $G = \langle a \rangle$, $o(a) = 24$, let $K = \langle a^{12} \rangle$ and $H = \langle a^6 \rangle$.
 - (a) In G/K , find the order of the elements Ka^2, Ka^3, Ka^4 , and Ka^5 .
 - (b) In G/H , find the order of the elements Ha^2, Ha^3, Ha^4 , and Ha^5 .
4. Let $G = \langle a \rangle \times \langle b \rangle$, where $o(a) = 8$ and $o(b) = 12$.
 - (a) If $K = \langle (a^2, b^3) \rangle$, find the order of $K(a^4, b)$ in G/K .
 - (b) If $K = \langle (a, b^2) \rangle$, find the order of $K(a^2, b)$ in G/K .

5. Let $G = D_{12} = \langle a, b \rangle$, where $o(a) = 6$, $o(b) = 2$, and $aba = b$.
 - (a) If $K = \langle a^2 \rangle$, find the order of Ka^2 , Ka^3 , Ka^5 , and Kba in G/K .
 - (b) If $K = \langle a^3, b \rangle$, find the order of Ka^2 , Ka^5 , and Kba^2 in G/K .
6. If Q denotes the quaternion group, show that $Q/Z(Q)$ has order 4. Is it cyclic or isomorphic to the Klein group? Support your answer.
7. Show that \mathbb{Q}/\mathbb{Z} is an infinite abelian group in which every element has finite order.
8. Let $K \subseteq H \subseteq G$ be finite groups, with $K \triangleleft G$. Show that $H/K = \{Kh \mid h \in H\}$ is a subgroup of G/K , and $|G/K : H/K| = |G : H|$.
9. If $K \triangleleft G$ and $o(g) = n$, $g \in G$, show that the order of Kg in G/K divides n .
10. If $K \triangleleft G$ has index m , show that $g^m \in K$ for all $g \in G$.
11. If $K \triangleleft G$ has index m and if $\gcd(m, n) = 1$, show that K contains every element of G of order n .
12. Let G be a finite group and let $K \triangleleft G$. If G/K has an element of order n , show that G has an element of order n .
13. Let $K \triangleleft G$. In each case, if both K and G/K have the given property, show that G also has the property.
 - (a) Trivial center.
 - (b) Every element has finite order.
 - (c) Every element has order a power of a fixed prime p .
 - (d) Finitely generated.
14. If $K \triangleleft G$ has prime index p , show that $G = K \cup Ka \cup \dots \cup Ka^{p-1}$ is a disjoint union for some $a \in G$.
15. If $G = \langle X \rangle$ is generated by X , and if $K \triangleleft G$, show that G/K is generated by $\{Kx \mid x \in X\}$.
16. Let H be a subset of G that is closed under the group operation. If $g^2 \in H$ for all $g \in G$, show that H is a normal subgroup of G and G/H is abelian.
17. If G is abelian, let $T(G)$ denote the set of elements in G of finite order.
 - (a) Show that $T(G)$ is a subgroup of G —the **torsion subgroup**.
 - (b) Call G a **torsion-free group** if $T(G) = \{1\}$. Show that $G/T(G)$ is torsion free.
 - (c) Call G a **torsion group** if $T(G) = G$. If H is a subgroup of G , show that G is a torsion group if and only if both H and G/H are torsion groups.
18. Let $K \subseteq H \subseteq G$ be groups, where $K \triangleleft G$ and $|G : K|$ is finite. Show that $|G/K : H/K|$ is also finite and that $|G/K : H/K| = |G : H|$.
19. Find G' in each case.
 - (a) G is abelian
 - (b) $G = Q$
 - (c) $G = D_6$
 - (d) $G = S_n$
20. Show that G' is a characteristic subgroup of G for every group G .
21. Show that $(G \times H)' = G' \times H'$.
22. If H is a subgroup of G , show that $H' \subseteq H \cap G'$. Show that this may not be equality.
23. Let $K \triangleleft G$.
 - (a) If $K \subseteq H$ where H is a subgroup of G , show that H/K is a subgroup of G/K .
 - (b) If \mathcal{X} is a subgroup of G/K , show that $\mathcal{X} = H/K$ where $H = \{h \in G \mid Kh \in \mathcal{X}\}$ is a subgroup of G containing K .
24. If $K \triangleleft G$ and $K \cap G' = \{1\}$, show that $K \subseteq Z(G)$ and that $Z[G/K] = Z(G)/K$.
25. Let $K \triangleleft G$.
 - (a) Show that $[Ka, Kb] = K[a, b]$ for all $a, b \in G$.
 - (b) If $K \subseteq G'$, show that $(G/K)' = G'/K$.
26. Let $K \subseteq Z(G)$ be a subgroup such that $G/K = \langle Kx_1, \dots, Kx_n \rangle$ where $x_i x_j = x_j x_i$ in G for all i and j . Show that G is abelian. (This extends Theorem 2.)

27. Let $K \subseteq H \subseteq G$ be groups with K characteristic in G . If H/K is characteristic in G/K , show that H is characteristic in G . [See Exercise 24 §2.8.]
28. (a) Show that $|G : Z(G)|$ cannot be a prime for any group G .
(b) Show that $G = D_4$ is a nonabelian group G such that $G/Z(G)$ is abelian.
29. If $k|n$, $k \geq 2$, show that D_n has a normal subgroup K such that $D_n/K \cong D_k$. [Hint: If D_n is generated by a and b where $o(a)=n$, $o(b)=2$, and $aba=b$, take $K=\langle a^k \rangle$.]
30. If $K = \{\varepsilon, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$, show that $S_4/K \cong D_3$. [Hint: Exercise 3 §2.8.]
31. Let $G = C_4 \times C_8$.
(a) Find subgroups H and K of G such that $H \cong K$ but $G/H \not\cong G/K$.
(b) Find subgroups P and Q of G such that $G/P \cong G/Q$ but $P \not\cong Q$.

2.10 THE ISOMORPHISM THEOREM

There is a connection between normal subgroups, homomorphisms and factor groups. The main relationship between these concepts is embodied in the isomorphism theorem, which is the principal result in this section and one of the most useful theorems in group theory. To describe it, we begin by identifying two subgroups associated with a group homomorphism $\alpha : G \rightarrow H$. The first is

The **image** of α , defined by $\text{im } \alpha = \alpha(G) = \{\alpha(g) \mid g \in G\}$.

This is a subgroup of H as was shown in Corollary 2 of Theorem 1 §2.5. We now turn to a subgroup of G determined by $\alpha : G \rightarrow H$:

The **kernel** of α , defined by $\ker \alpha = \{k \in G \mid \alpha(k) = 1\}$.

Theorem 1. *Let $\alpha : G \rightarrow H$ be a group homomorphism.*

- (1) $\alpha(G)$ is a subgroup of H .
- (2) $\ker \alpha$ is a normal subgroup of G .

Proof. (1) This is Corollary 2 of Theorem 1 §2.5.

(2) We have $1 \in \ker \alpha$ because $\alpha(1) = 1$. If $k, k' \in \ker \alpha$, then

$$\alpha(kk') = \alpha(k) \cdot \alpha(k') = 1 \cdot 1 = 1 \quad \text{and} \quad \alpha(k^{-1}) = \alpha(k)^{-1} = 1^{-1} = 1.$$

Hence $kk' \in \ker \alpha$ and $k^{-1} \in \ker \alpha$, so $\ker \alpha$ is a subgroup. If $g \in G$ and $k \in K$ then

$$\alpha(gkg^{-1}) = \alpha(g) \cdot \alpha(k) \cdot \alpha(g^{-1}) = \alpha(g) \cdot 1 \cdot \alpha(g)^{-1} = 1.$$

This shows that $g(\ker \alpha)g^{-1} \subseteq \ker \alpha$ for all $g \in G$, and so proves that $\ker \alpha \triangleleft G$. ■

Note that the image of a homomorphism $\alpha : G \rightarrow H$ need not be normal in H . For example, if K is any subgroup of H , define the **inclusion mapping** $\iota : K \rightarrow H$ by $\iota(k) = k$ for all $k \in K$. This is a one-to-one homomorphism, but $\iota(K) = K$ need not be normal in H .

Theorem 1 shows that kernels of homomorphisms from G are normal in G . Conversely, every normal subgroup of a group G arises as the kernel of some homomorphism with G as domain:

Theorem 2. *If $K \triangleleft G$, then $K = \ker \varphi$ where $\varphi : G \rightarrow G/K$ is the coset mapping.*

Proof. The coset map φ is defined by $\varphi(g) = Kg$ for all $g \in G$ and is a homomorphism by Theorem 1 §2.9. Because K is the unity of the group G/K , we have $g \in \ker \varphi$ if and only if $Kg = K$, if and only if $g \in K$. Hence $\ker \varphi = K$. ■

Many important subgroups are kernels of naturally occurring homomorphisms; indeed, the easiest way to verify that a subgroup of a group G is normal in G is often to exhibit it as the kernel of a homomorphism with G as domain.

Example 1. The absolute value homomorphism $\mathbb{C}^* \rightarrow \mathbb{R}^+$ given by $z \mapsto |z|$ has kernel the circle group $\mathbb{C}^0 = \{z \in \mathbb{C}^* \mid |z| = 1\}$.

Example 2. The kernel of the determinant homomorphism $A \mapsto \det A$ from $GL_n(\mathbb{R}) \rightarrow \mathbb{R}^*$ is the **special linear group** $SL_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}) \mid \det A = 1\}$.

Example 3. If G is a group and $g \in G$ has finite order n , let $\alpha : \mathbb{Z} \rightarrow G$ be the exponent mapping given by $\alpha(k) = g^k$. Then $\ker \alpha = n\mathbb{Z}$ by Theorem 2 §2.4.

Example 4. Show that $A_n \triangleleft S_n$ by exhibiting A_n as a kernel.

Solution. Define the **sign** of a permutation $\sigma \in S_n$ by $\operatorname{sgn} \sigma = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd} \end{cases}$.

Then the sign mapping $\alpha : S_n \rightarrow \{1, -1\}$ given by $\alpha(\sigma) = \operatorname{sgn} \sigma$ is a homomorphism (see Exercise 29 §1.4). Clearly $\ker \alpha = A_n$. □

Example 5. The **trivial** homomorphism $G \rightarrow H$ is the only one with G as kernel.

It is clear that a homomorphism $\alpha : G \rightarrow H$ is onto if and only if $\alpha(G) = H$, that is, if and only if the image $\alpha(G)$ is as large a subgroup of H as possible. The next theorem shows that α is one-to-one if and only if $\ker \alpha$ is as small as possible.

Theorem 3. If $\alpha : G \rightarrow H$ is a homomorphism, then α is one-to-one if and only if $\ker \alpha = \{1\}$.

Proof. If α is one-to-one, let $g \in \ker \alpha$. Thus $\alpha(g) = 1 = \alpha(1)$, so $g = 1$ because α is one-to-one. Hence $\ker \alpha = \{1\}$. Conversely, let $\ker \alpha = \{1\}$ and suppose that $\alpha(a) = \alpha(b)$ where a and b are in G . Then $\alpha(ab^{-1}) = \alpha(a)\alpha(b)^{-1} = 1$, so $ab^{-1} \in \ker \alpha = \{1\}$. This shows that $ab^{-1} = 1$ and hence that $a = b$. Thus α is one-to-one. ■

Theorem 3 is used frequently to test when a homomorphism is one-to-one.

We now come to one of the most useful theorems in group theory.

Theorem 4. Isomorphism Theorem³⁵. Let $\alpha : G \rightarrow H$ be a group homomorphism and write $K = \ker \alpha$. Then

$$\alpha(G) \cong G/\ker \alpha.$$

Proof. Write $K = \ker \alpha$ for simplicity, and define

$$\bar{\alpha} : G/K \rightarrow \alpha(G) \text{ by } \bar{\alpha}(Kg) = \alpha(g) \text{ for all } Kg \in G/K.$$

First $\bar{\alpha}$ is well defined; that is, $Kg = Kg_1$ implies that $\alpha(g) = \alpha(g_1)$. In fact,

$$Kg = Kg_1 \Leftrightarrow gg_1^{-1} \in K \Leftrightarrow \alpha(gg_1^{-1}) = 1 \Leftrightarrow \alpha(g) = \alpha(g_1).$$

³⁵This result goes back to Camille Jordan (1838–1922) in his book *Traité des Substitutions* (1870), where the concept of a homomorphism was introduced.

Hence $\bar{\alpha}$ is well defined (\Rightarrow) and one-to-one (\Leftarrow). As $\bar{\alpha}$ is clearly onto $\alpha(G)$, it remains to show that it is a homomorphism. But

$$\bar{\alpha}(Kg K g_1) = \bar{\alpha}(Kgg_1) = \alpha(gg_1) = \alpha(g) \cdot \alpha(g_1) = \bar{\alpha}(Kg) \cdot \bar{\alpha}(Kg_1)$$

holds for all Kg and Kg_1 in G/K . \blacksquare

If G is a group, a group of the form $\alpha(G)$ where $\alpha : G \rightarrow H$ is some homomorphism is called a **homomorphic image** of G . Hence the isomorphism theorem shows that the factor groups of G and the homomorphic images of G are the same set of groups up to isomorphism.

Remark. The diagram to the right depicts the mappings α and $\bar{\alpha}$ in the isomorphism theorem. Here $K = \ker \alpha$ as in the theorem, and the mapping $\varphi : G \rightarrow G/K$ is the coset mapping. Note that $\alpha = \bar{\alpha}\varphi$ is a factorization of the (arbitrary) homomorphism α as a composite where φ is onto and $\bar{\alpha}$ is one-to-one. Indeed, $\bar{\alpha}\varphi(g) = \bar{\alpha}[\varphi(g)] = \bar{\alpha}(Kg) = \alpha(g)$ for all $g \in G$. Moreover, $\bar{\alpha}$ is the *only* homomorphism $G/K \rightarrow H$ with the property that $\bar{\alpha}\varphi = \alpha$. Indeed, if this condition holds then the action of $\bar{\alpha}$ is determined: $\bar{\alpha}(Kg) = \bar{\alpha}[\varphi(g)] = \bar{\alpha}\varphi(g) = \alpha(g)$ for all Kg in G/K . Hence:

$$\begin{array}{ccc} G & \xrightarrow{\alpha} & H \\ \varphi \downarrow & \nearrow \bar{\alpha} & \\ G/K & & \end{array}$$

Corollary. Let $\alpha : G \rightarrow H$ be a group homomorphism. Then α factors uniquely as $\alpha = \bar{\alpha}\varphi$ where $\varphi : G \rightarrow G/\ker \alpha$ is the coset map, and $\bar{\alpha} : G/\ker \alpha \rightarrow H$ is defined in Theorem 4. Note that φ is onto, and $\bar{\alpha}$ is one-to-one.

The isomorphism theorem is a marvelous result. It sheds light on nearly every situation to which it is applied. It is used as follows: If we want to show that $G/K \cong H$, we find an onto homomorphism $G \rightarrow H$ with kernel K . As a bonus, the fact that K is a kernel automatically proves that it is normal in G . Examples 6–9 illustrate the use of the isomorphism theorem.

Example 6. If G is a cyclic group, show that $G \cong \mathbb{Z}$ or $G \cong \mathbb{Z}_n$.

Solution. Let $G = \langle a \rangle$ and define $\alpha : \mathbb{Z} \rightarrow G$ by $\alpha(k) = a^k$ for all $k \in \mathbb{Z}$. This is an onto homomorphism and $\ker \alpha = \{k \mid a^k = 1\}$. If $o(a)$ is infinite, $\ker \alpha = \{0\}$ and the isomorphism theorem gives $G \cong \mathbb{Z}/\{0\} \cong \mathbb{Z}$. If $o(a) = n$, then $\ker \alpha = n\mathbb{Z}$ and $G \cong \mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$. \square

Example 7. Let $K \triangleleft G$ and $K_1 \triangleleft G_1$. Show that $(K \times K_1) \triangleleft (G \times G_1)$ and

$$(G \times G_1)/(K \times K_1) \cong (G/K) \times (G_1/K_1).$$

Solution. We define $\alpha : (G \times G_1) \rightarrow (G/K) \times (G_1/K_1)$ by $\alpha(g, g_1) = (Kg, K_1g_1)$. It is routine to verify that this is an onto homomorphism, and $\ker \alpha = K \times K_1$. The isomorphism theorem now gives all our assertions. \square

Example 8. Show that $\mathbb{R}/\mathbb{Z} \cong \mathbb{C}^0$ where $\mathbb{C}^0 = \{e^{i\theta} \mid \theta \in \mathbb{R}\}$ is the circle group.

Solution. We define $\alpha : \mathbb{R} \rightarrow \mathbb{C}^0$ by $\alpha(x) = e^{2\pi xi}$. We have

$$\alpha(x+y) = e^{2\pi(x+y)i} = e^{2\pi xi}e^{2\pi yi} = \alpha(x) \cdot \alpha(y)$$

so α is a homomorphism. It is clearly onto, and

$$\alpha(x) = 1 \Leftrightarrow e^{2\pi xi} = 1 \Leftrightarrow x \in \mathbb{Z}.$$

Thus, $\ker \alpha = \mathbb{Z}$ and the isomorphism theorem does the rest. \square

If we are interested in determining *all* homomorphisms $\alpha : G \rightarrow G_1$, the fact that $\alpha(G_1)$ is isomorphic to $G / (\ker \alpha)$ is useful because sometimes we can determine the normal subgroups of G . In Example 9 §2.5, we showed that there are at most six homomorphisms: $S_3 \rightarrow C_6$, and hence at most 6 from $D_3 \rightarrow C_6$. Using the isomorphism theorem, we can show that in fact there are only two.

Example 9. Write $D_3 = \{1, a, a^2, b, ba, ba^2\}$, where $o(a) = 3$, $o(b) = 2$, and $aba = b$, and write $C_6 = \langle c \rangle$, where $o(c) = 6$. Show that there are only two homomorphisms, $D_3 \rightarrow C_6$, the trivial one and

$$\alpha : D_3 \rightarrow C_6 \text{ defined by } \alpha(b^k a^m) = c^{3k} \text{ for all } b^k a^m \in D_3.$$

Solution. We know that D_3 has only three normal subgroups: $\{1\}$, D_3 , and $K = \langle a \rangle$. Thus if $\alpha : D_3 \rightarrow C_6$ is a homomorphism, $\ker \alpha$ must be one of them. It is impossible that $\ker \alpha = \{1\}$ because then $\alpha(D_3) \cong D_3$ would be a nonabelian subgroup of C_6 . If $\ker \alpha = D_3$ then α is the trivial homomorphism. So assume that $\ker \alpha = K = \langle a \rangle$. In this case, let $\varphi : D_3 \rightarrow D_3/K$ be the coset map. If α exists, the corollary to the isomorphism theorem guarantees that $\alpha = \sigma \varphi$, where $\sigma : D_3/K \rightarrow \alpha(D_3)$ is an isomorphism. In this case $D_3/K = \{K, bK\}$ is cyclic of order 2, so $\alpha(D_3)$ is the (unique) subgroup of order 2 in C_6 ; that is $\alpha(D_3) = \{1, c^3\}$. Clearly, $\sigma(K) = 1$ and $\sigma(bK) = c^3$, so $\alpha = \sigma \varphi$ is given by

$$\begin{array}{ccc} D_3 & \xrightarrow{\alpha} & \alpha(D_3) \\ \varphi \downarrow & \nearrow \sigma & \\ D_3/K & & \end{array}$$

$$\alpha(b^k a^m) = \sigma \varphi(b^k a^m) = \sigma(b^k a^m K) = \sigma(bK)^k \cdot \sigma(aK)^m = (c^3)^k \cdot 1^m = c^{3k}. \quad \square$$

We conclude this section with one more result using the isomorphism theorem. Recall (Example 18 §2.5) that the set $\text{inn } G$ of all inner automorphisms of a group G is a subgroup of the group $\text{aut } G$ of all automorphisms of G .

Theorem 5. If G is any group, then $G/Z(G) \cong \text{inn } G$.

Proof. If $a \in G$, recall that the inner automorphism $\sigma_a : G \rightarrow G$ is defined by $\sigma_a(g) = aga^{-1}$ for all $g \in G$. Then $\sigma_a \sigma_b = \sigma_{ab}$ for all $a, b \in G$ (Example 17 §2.5), and so $\theta(a) = \sigma_a$ defines a group homomorphism $\theta : G \rightarrow \text{aut } G$. Clearly, $\theta(G) = \text{inn } G$, and

$$\ker \theta = \{a \in G \mid \sigma_a = 1_G\} = \{a \in G \mid aga^{-1} = g \text{ for all } g \in G\} = Z(G).$$

The result now follows from the isomorphism theorem. \blacksquare

Example 10. Show that $\text{inn } S_3 \cong S_3$. Show further that $\text{inn } S_3 = \text{aut } S_3$.

Solution. $Z(S_3) = \{\varepsilon\}$ is easily verified, so $S_3 \cong \text{inn } S_3$ by Theorem 5. Hence $|\text{inn } S_3| = 6$ so, since $\text{inn } S_3 \subseteq \text{aut } S_3$, it suffices to show that $|\text{aut } S_3| \leq 6$. But $S_3 = \{1, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$ where $o(\sigma) = 3$ and $o(\tau) = 2$. So if $\theta : S_3 \rightarrow S_3$ is an

automorphism then $o(\theta(\sigma)) = o(\sigma) = 3$, so $\theta(\sigma) = \sigma$ or σ^2 . Similarly $\theta(\tau) = \tau, \tau\sigma$ or $\tau\sigma^2$, so there are at most $2 \cdot 3 = 6$ choices for θ because $S_3 = \langle \sigma, \tau \rangle$. \square

Exercises 2.10

1. Let $G = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \mid a, b, c \in \mathbb{R}, a \neq 0, c \neq 0 \right\}$ and $K = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \mid b \in \mathbb{R} \right\}$, subgroups of $\text{GL}_n(\mathbb{R})$. Show that $K \triangleleft G$ by exhibiting K as the kernel of a group homomorphism $G \rightarrow \mathbb{R}^* \times \mathbb{R}^*$.
2. Show that the following are equivalent for a group homomorphism $\alpha : G \rightarrow G_1$.
 - (a) α is trivial
 - (b) $\ker \alpha = G$
 - (c) $\alpha(G) = \{1\}$
3. Let H be a subgroup of G with $|G : H| = 2$, and define $\alpha : G \rightarrow \{1, -1\}$ by $\alpha(a) = \begin{cases} 1, & \text{if } a \in H \\ -1, & \text{if } a \notin H \end{cases}$. Show that α is a homomorphism and that $\ker \alpha = H$.
4. If $\alpha : G \rightarrow G_1$ is a group homomorphism and if X is a subgroup of $\alpha(G)$, the preimage of X under α is defined by $\alpha^{-1}(X) = \{g \in G \mid \alpha(g) \in X\}$. For example $\alpha^{-1}(\{1\}) = \ker \alpha$. [Note: The notation α^{-1} here is *not* intended to imply that α is an isomorphism.]
 - (a) Show that $\alpha^{-1}(X)$ is a subgroup of G , normal if $X \triangleleft \alpha(G)$.
 - (b) Show that $X \subseteq Y$ if and only if $\alpha^{-1}(X) \subseteq \alpha^{-1}(Y)$.
 - (c) Show that $\alpha^{-1}(X \cap Y) = \alpha^{-1}(X) \cap \alpha^{-1}(Y)$.
5. Let $\rho_m : G \rightarrow G$ be the m -power map: $\rho_m(g) = g^m$. Assume G is abelian and $|G| = n$.
 - (a) Show that $\ker \rho_m = \{g \mid g^d = 1\}$ where $d = \gcd(m, n)$.
 - (b) If m and n are relatively prime, show that ρ_m is an automorphism.
 - (c) If $G = \langle a \rangle$ is cyclic, show that every automorphism of G arises as in (b).
6. Let $\alpha : G \rightarrow G_1$ be a group homomorphism with $\ker \alpha = K$. For $a \in G$, show that $Ka = \{g \in G \mid \alpha(g) = \alpha(a)\}$.
7. If $\alpha : G \rightarrow G_1$ is a group homomorphism and both $\alpha(G)$ and $\ker \alpha$ are finitely generated, show that G is finitely generated.
8. Find all group homomorphisms
 - (a) $C_6 \rightarrow K_4$
 - (b) $C_3 \rightarrow A_4$
 - (c) $D_3 \rightarrow C_4$
 - (d) $A_4 \rightarrow C_3$
9. If $K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$, is there a group homomorphism $\alpha : S_4 \rightarrow A_4$, with $\ker \alpha = K$? Support your answer.
10. Determine if there exists an onto group homomorphism in each case:
 - (a) $\alpha : S_3 \rightarrow K_4$
 - (b) $\alpha : S_3 \rightarrow C_3$
 - (c) $\alpha : S_3 \rightarrow C_2$
 - (d) $\alpha : \mathbb{R}^* \rightarrow C_2$
11. If G is a group, let $\theta : G \rightarrow G \times G$ be defined by $\theta(g) = (g, g)$.
 - (a) Show that θ is a one-to-one group homomorphism.
 - (b) Show that the following conditions are equivalent: (1) G is abelian; (2) $\theta(G)$ is normal in $G \times G$; and (3) $\theta(G) = \ker \varphi$ for some homomorphism $\varphi : G \times G \rightarrow G$.
12. Show that a group G is simple if and only if every nontrivial group homomorphism $G \rightarrow G_1$ is one-to-one.
13. If G is a simple group, show that there is a nontrivial group homomorphism $G \rightarrow G_1$ if and only if G_1 has a subgroup isomorphic to G .
14. If n is odd, show that there are at most 36 group homomorphisms $D_n \rightarrow A_4$.
15. If $|G| \geq 2$ and $\text{aut } G$ is cyclic, show that G is abelian and that $\text{aut } G$ is finite and of even order. [Hint: Theorem 2 §2.9 and Theorem 5.]

16. If $\text{aut } G$ is simple, show that G is abelian or $G/Z(G)$ is simple. [Hint: Exercise 16 §2.8.]

17. Let $\alpha : G \rightarrow G_1$ be a group homomorphism, as shown in the figure at the right.

(a) Show that $\alpha(G') \subseteq G'_1$. [Hint: Show $\alpha([a, b]) = [\alpha(a), \alpha(b)]$ for all $a, b \in G$.]

(b) If $\varphi : G \rightarrow G/G'$ and $\varphi_1 : G_1 \rightarrow G_1/G'_1$ are the coset maps, show that a unique homomorphism $\bar{\alpha} : G/G' \rightarrow G_1/G'_1$ exists such that $\bar{\alpha}\varphi = \varphi_1\alpha$ (see the diagram).

$$\begin{array}{ccc} G & \xrightarrow{\alpha} & G_1 \\ \varphi \downarrow & & \downarrow \varphi_1 \\ G/G' & \xrightarrow{\bar{\alpha}} & G_1/G'_1 \end{array}$$

18. If $G = H \times K$ and $K_1 = \{(1, k) \mid k \in K\}$, show that $K_1 \triangleleft G$, $K_1 \cong K$ and $G/K_1 \cong H$.

19. Let $G = GL_n(\mathbb{R})$, let $K = \{A \mid \det A = 1\}$ and let $K_1 = \{A \mid \det A = \pm 1\}$.

(a) Show $K \triangleleft G$ and $G/K \cong \mathbb{R}^*$. (b) Show $K_1 \triangleleft G$ and $G/K_1 \cong \mathbb{R}^+$.

20. Let $G = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \mid a, b, c \in \mathbb{R}; a \neq 0, c \neq 0 \right\}$. If $K = \left\{ \begin{bmatrix} 1 & b \\ 0 & 1 \end{bmatrix} \mid b \in \mathbb{R} \right\}$, show that $K \triangleleft G$ and $G/K \cong \mathbb{R}^* \times \mathbb{R}^*$.

21. Show that $\mathbb{C}^*/\mathbb{C}^0 \cong \mathbb{R}^+$, where $\mathbb{C}^0 = \{z \mid |z| = 1\}$ is the circle group.

22. Show that $\mathbb{R}^*/\{1, -1\} \cong \mathbb{R}^+$.

23. If $a, b \in \mathbb{R}$, define $\tau_{a,b} : \mathbb{R} \rightarrow \mathbb{R}$ by $\tau_{a,b}(x) = ax + b$ for all $x \in \mathbb{R}$. It follows that $G = \{\tau_{a,b} \mid a, b \in \mathbb{R}; a \neq 0\}$ is a subgroup of $S_{\mathbb{R}}$. Show that $K = \{\tau_{1,b} \mid b \in \mathbb{R}\}$ is a normal subgroup of G and $G/K \cong \mathbb{R}^*$.

24. Consider $M_2(\mathbb{Z})$ as a group under addition. For $n \geq 2$, show that $M_2(n\mathbb{Z}) \triangleleft M_2(\mathbb{Z})$ and $M_2(\mathbb{Z})/M_2(n\mathbb{Z}) \cong M_2(\mathbb{Z}_n)$ —all additive groups.

25. If G is abelian, let $K = \{(g, g, g) \mid g \in G\}$. Show that $K \triangleleft G \times G \times G$ and $G \times G \times G/K \cong G \times G$.

26. If $G/K \cong H$, show that there is an onto homomorphism $\alpha : G \rightarrow H$ with $\ker \alpha = K$.

27. If $\alpha : G \rightarrow G_1$ is a group homomorphism and $K \triangleleft G$ with $\ker \alpha \subseteq K$, show that $\alpha(K) \triangleleft \alpha(G)$ and $\alpha(G)/\alpha(K) \cong G/K$.

28. Let G be a finite abelian group. Show that the following conditions are equivalent for an integer m : (1) $g^m = 1$ in G implies that $g = 1$; and (2) every element $g \in G$ has an m th root, that is, $g = a^m$ for some $a \in G$. Compare your results with those of Exercise 16 §2.6.

29. Let $G = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} \mid a, b, c \in \mathbb{R} \right\}$.

(a) Show that G is a subgroup of $M_3(\mathbb{R})^*$ and that $Z(G) \cong \mathbb{R}$.

(b) Show that $G/Z(G) \cong \mathbb{R} \times \mathbb{R}$.

30. Use the isomorphism theorem to show that, if $m|n$, then $\mathbb{Z}_n/\langle \bar{m} \rangle \cong \mathbb{Z}_m$.

31. Let $s = \text{lcm}(m, n)$. Show that \mathbb{Z}_s is isomorphic to a subgroup of $\mathbb{Z}_m \times \mathbb{Z}_n$. [Hint: Think of \mathbb{Z}_m as $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$.]

32. Show that every infinite homomorphic image of \mathbb{Z} is isomorphic to \mathbb{Z} .

33. Describe the homomorphic images of each group.

(a) \mathbb{Z}_4 (b) A simple group G (c) A_4 [Hint: Exercise 3 §2.8.]

34. If $|G| \geq 3$, show that G has at least two automorphisms. [Hint: Theorem 5.]

35. Let $\alpha : G \rightarrow G_1$ be an onto group homomorphism. If X is a subgroup of G_1 , define $\alpha^{-1}(X) = \{g \in G \mid \alpha(g) \in X\}$ as in Exercise 4. If $X \triangleleft G_1$ show that $\alpha^{-1}(X) \triangleleft G$ and $G/\alpha^{-1}(X) \cong G_1/X$.

- 36.** If X and Y are additive abelian groups, let $\text{hom}(X, Y)$ denote the set of all group homomorphisms $\alpha : X \rightarrow Y$. If $\alpha, \beta \in \text{hom}(X, Y)$, define $\alpha + \beta : X \rightarrow Y$ by $(\alpha + \beta)(x) = \alpha(x) + \beta(x)$ for all $x \in X$.
- Show that $\text{hom}(X, Y)$ is an abelian group under this addition.
 - Show that $Y \cong \text{hom}(\mathbb{Z}, Y)$ for every additive abelian group Y .
 - Show that $\text{hom}(\mathbb{Z}_m, \mathbb{Z}_n) \cong \mathbb{Z}_d$, where $d = \gcd(m, n)$. [Hint: If $e = n/d$, define $\alpha_k : \mathbb{Z}_m \rightarrow \mathbb{Z}_n$ by $\alpha_k(\tilde{x}) = k\tilde{x}$, where $\tilde{x} = x + m\mathbb{Z} \in \mathbb{Z}_m$ and $\tilde{x} = x + n\mathbb{Z} \in \mathbb{Z}_n$.]
- 37.** If G is a group and $g_i \in G$ for all $i \geq 0$, let $[g_i] = (g_0, g_1, g_2, \dots)$ denote an infinite sequence from G . Define $[g_i] = [h_i]$ if and only if $g_i = h_i$ for all $i \geq 0$ and define $[g_i] \cdot [h_i] = [g_i h_i]$. Write $G^\omega = \{[g_i] \mid g_i \in G\}$.
- Show that G^ω is a group with the preceding multiplication.
 - Show that $G_0 = \{[g_i] \mid g_0 \in G, g_i = 1 \text{ for all } i \geq 1\}$ is a normal subgroup of G^ω , and $G^\omega/G_0 \cong G^\omega$.
 - Let F denote the set of mappings $\mathbb{N} \rightarrow G$ and, if $f, g \in F$, define $fg \in F$ by $fg(i) = f(i) \cdot g(i)$ for all $i \in \mathbb{N}$. Show that F is a group. What is the relationship between F and G^ω ? Support your answer.
- 38.** If $K \triangleleft G$ show that $C(K) \triangleleft G$ and $G/[C(K)]$ is isomorphic to a subgroup of $\text{aut } K$, where $C(K) = \{a \in G \mid ak = ka \text{ for all } k \in K\}$. [Hint: Theorem 5.]

2.11 AN APPLICATION TO BINARY LINEAR CODES

The value of mathematics in any science lies more in disciplined analysis and abstract thinking than in particular theories or techniques.

—Alan Tucker

Coding theory is concerned with the transmission of information over a *channel* that is affected by *noise*. The noise causes errors, and the general aim is to detect such errors when they occur and to correct them if possible. Such codes are used every day in communication systems such as radio, television, and telephone; in data storage systems such as those used by banks; in the internal circuits of computers; and in many other systems where information is being processed. With the advent of computers, information is often expressed in *digital* form, that is as strings of 0s and 1s which computers can easily handle. Consequently we deal with *binary* codes that are based on $\mathbb{Z}_2 = \{0, 1\}$.

General coding theory originated in the 1940s, primarily with the work of Claude Shannon. He created a mathematical theory of information and proved that certain codes exist which can transmit information at near optimal rates with arbitrarily small chance of error. In 1950, Richard W. Hamming discovered the error-detecting and error-correcting codes that now bear his name. Many of these codes are widely used today.

Example 1 concretely illustrates many of the features of general coding.

Example 1. Suppose that a spacecraft is orbiting the moon, and assume that the message 1 or 0 is to be sent instructing the mission commander to land or not.

Because of static interference (noise) the probability³⁶ is 0.1 that an error will occur during transmission (and hence a probability of 0.9 that no error will occur). To ensure accuracy, the earth station transmits five signals: 11111 instead of 1 and 00000 for 0. The spacecraft computer receives a five-digit message and decodes it by a simple majority: It concludes that 11111 was sent if more 1s than 0s are received and that 00000 was sent otherwise. For example, if it receives 11001 it concludes that 11111 was sent. Thus, the spacecraft computer will get the wrong message if and only if three or more errors occur in transmission and (assuming successive errors occur independently) the probability of this happening³⁷ is 0.00856. This probability is less than 1%, even though there is a 10% chance of error on any one transmission. This decision method is called **maximum likelihood decoding**. □

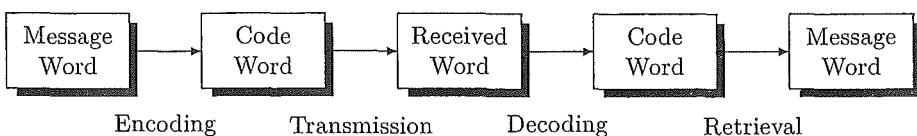
Example 1 is a good illustration of the way coding works. A sender has a message to send (say, 1 in Example 1). It is encoded (as 11111) and transmitted over a noisy channel where it is received (as, say, 11001) and decoded (as 1) before being sent to the receiver. In Example 1, the coding process can detect errors and correct them with a probability of less than 0.01 of being wrong.

In general, it is desirable to have more messages than 1 or 0 available for encoding and transmission. For convenience (and due to the ubiquity of computers) we assume that our messages, and the encoded messages to be transmitted, are strings of 0s and 1s. We use the following notation. If $n \geq 1$, let

$$B^n = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \cdots \times \mathbb{Z}_2$$

denote the direct product of n copies of the (additive abelian) group $\mathbb{Z}_2 = \{0, 1\}$. The elements of B^n are called **words of length n** and, for convenience, we write them as strings of 0s and 1s rather than as n -tuples. Thus, 110101 in B^6 stands for (1, 1, 0, 1, 0, 1). We call the individual 0s and 1s the **bits** of the word (an abbreviation for *binary digits*). A subset C of B^n , with $|C| \geq 2$, is called an n -binary code (or simply an n -code). The words in C are called **code words**.

We describe the general coding process in the diagram.



A set of words, called **message words**, is given in B^k . They are paired with a set C of longer words in B^n , $n \geq k$, which will actually be transmitted. Thus C is

³⁶We treat probability informally here. The probability that an event occurs is the long-term proportion of the time that the event does indeed occur. Thus probabilities are numbers between 0 and 1. A probability of 0 means that the event in question is impossible; a probability of 1 means that the event is certain to occur; and a probability of 0.5 means that the event is as likely as not to occur.

³⁷The probability is computed as $\binom{5}{3}(0.1)^3(0.9)^2 + \binom{5}{4}(0.1)^4(0.9)^1 + \binom{5}{5}(0.1)^5 = 0.00856$. It is based on the assumption that at most one error occurs in each digit transmitted and that these errors occur independently.

an n -code, and the process of passing from a message to the corresponding code word is called **encoding**. Only code words are transmitted but, as some bits may be altered during transmission, words other than code words may be received. The sole purpose of the encoding process is to enable the receiver to detect errors and, if there are not too many, to correct them. The encoding and transmission processes are usually quite simple. The message words in B^k are paired with code words in B^n in such a way that passing back and forth is easy. A common method is to add extra bits (called **check bits**) to the end of the message so that the message itself forms the first k bits of the code word (making retrieval easy). The transmission process is more complex, and the design of codes that are easy and inexpensive to transmit (using, say, shift registers) is an important problem that we do not consider here. The most mathematically interesting part of the process is decoding. A method must be devised to detect bit errors in the received word and, hopefully, to correct them and so reconstruct the transmitted code word. The transmission and decoding part of the process begins and ends with code words, so we concentrate on constructing codes and pay less attention to encoding and retrieving.

In Example 1, the 5-code $\{00000, 11111\}$ has so few code words that a system (majority rule) of decoding can correct errors with a small probability of error. However, sometimes (for example, when retransmission is easy and inexpensive) all that is needed is to detect errors. Example 2 gives one such system that is commonly used.

Example 2. Parity-check Codes are n -codes that are constructed as follows. The message words are the elements of B^{n-1} , and we form the code words by adding one extra bit at the end, selecting it so that the total number of 1s is even (equivalently, the sum of the bits (in \mathbb{Z}_2) is 0). Such words are said to have even parity. Thus, the 4-parity-check code C is

Message words (B^3): 000 001 010 011 100 101 110 111

Code words (C): 0000 0011 0101 0110 1001 1010 1100 1111

If a member of C is transmitted and one error occurs, the received word will have an odd number of 1s (**odd parity**) and so the error is detected. This code can thus detect any odd number of errors, but it cannot detect an even number of errors and it cannot correct any errors. Nonetheless, it is used in banking (the last digit of an account number is often a control digit) and in the internal arithmetic of digital computers. \square

Nearest Neighbor Decoding

Many important error-correcting codes operate in the following way. A method is found to define the distance between two words in B^n . Then a code $C \subseteq B^n$ is found whose members are so far apart that, if any one bit (say) in a code word c is changed, the new word w is still closer to c than to any other word in the code. Thus, if c is transmitted and one error occurs, the received word w can be corrected

by replacing it with the code word closest to it. We state this more compactly as follows.

Nearest Neighbor Decoding *Let C be an n -code. If a word w is received, it is decoded as the code word in C closest to it. (If more than one candidate appears, choose arbitrarily³⁸.)*

Codes can be constructed that will correct any finite number of errors using nearest neighbor decoding.

Of course the whole thing depends on the existence of an appropriate distance function on B^n . If a word c is transmitted and t errors occur, the received word w will differ from c in exactly t bits. This is the distance between c and w .

More precisely, let v and w be words in B^n . The **Hamming distance**³⁹ $d(v, w)$ between v and w is the number of coordinates at which their corresponding bits differ. Thus, if $v = v_1 v_2 \dots v_n$ and $w = w_1 w_2 \dots w_n$, where the v_i and w_i are the bits, then $d(v, w)$ is the number of indices i such that $v_i \neq w_i$. Define the **Hamming weight** of w by $\text{wt } w = d(w, 0)$. Thus, $\text{wt } w$ is the number of 1s occurring as bits of the word w .

The following theorem gives some fundamental properties of the Hamming weight and distance functions. The proof uses the fact that B^n is an additive group under componentwise operations. Thus two words are added by adding corresponding bits modulo 2. For example,

$$10101 + 11011 = 01110 \quad \text{in } B^5.$$

Note that the unity is the word $000 \dots 0$, each of whose bits is 0, which we denote 0. Also, $-w = w$ for each word w in B^n , but we write $v - w$ for clarity.

Theorem 1. *Let u , v , and w be words in B^n .*

- (1) $d(v, w) = \text{wt}(v - w)$.
- (2) $d(v, w) = d(w, v)$.
- (3) $d(v, w) = 0$ if and only if $v = w$.
- (4) $d(u, w) \leq d(u, v) + d(v, w)$.

Proof. (1) A bit of $v - w$ is a 1 if and only if v and w differ at that coordinate. Hence the number of bits of $v - w$ that are 1s equals the number of coordinates where v and w differ. This is (1).

(2), (3). We leave the proofs to the reader.

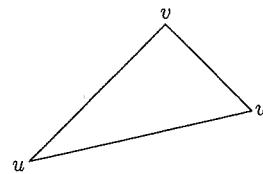
(4) Write $x = u - v$ and $y = v - w$, so that $x + y = u - w$. Then, using (1), condition (4) becomes $\text{wt}(x + y) \leq \text{wt } x + \text{wt } y$. Now let x_i and y_i denote the i th bits of x and y , respectively. Then the $\text{wt}(x + y)$ is the number of values of i for which $x_i + y_i = 1$. Hence (4) certainly holds if $x_i + y_i = 1$ implies that $x_i = 1$ or $y_i = 1$. But this implication is clear because $x_i = 0 = y_i$ implies that $x_i + y_i = 0$. ■

Properties (2), (3), and (4) of Theorem 1 justify calling d a distance function on B^n . The first two are clearly true of ordinary distance. With respect to

³⁸If it is feasible, retransmission may be called for in this case.

³⁹The name honors Richard W. Hamming. Distance functions are also called metrics.

property (4), we may regard u, v , and w as the vertices of a triangle (see the figure). Then (4) asserts that the length of one side of a triangle is not greater than the sum of the lengths of the other two sides. For this reason we call (4) the **triangle inequality**.



This geometric terminology for Hamming distance is useful for discussing nearest neighbor decoding. If w is a word in B^n and $r \geq 0$ is a real number, the set

$$S_r(w) = \{v \in B^n \mid d(v, w) \leq r\}$$

is called the **ball of radius r about w** or simply the r -ball about w . We use this to describe how to construct a code C that can detect (or correct) t errors.

Suppose that a code word c is transmitted and a word w is received with s errors, where $1 \leq s \leq t$. Then s is the number of coordinates at which the digits of c and w differ; that is, $s = d(c, w)$. Hence $S_t(c)$ consists of all possible received words where at most t errors have occurred. We first assume that C has the property that no code word lies in the t -ball of another code word. Because $w \in S_t(c)$ and $w \neq c$, this means that w is not a code word and that the error has been detected. If we strengthen the assumption on C to require that the t -balls about code words are pairwise disjoint, then w belongs to a unique ball (that about c), so w will be correctly decoded as c .

To describe when this happens, let C be an n -code. The **minimum distance** d of C is defined to be the smallest distance between two distinct code words in C . That is,

$$d = \min\{d(v, w) \mid v, w \in C; v \neq w\}.$$

Theorem 2. Let C be an n -code with minimum distance d . Assume that nearest neighbor decoding is used.

- (1) If $t + 1 \leq d$, then C can detect⁴⁰ t errors.
- (2) If $2t + 1 \leq d$, then C can correct t errors.

Proof. (1) If $c \in C$, the t -ball $S_t(c)$ contains no other code word because $t < d$. Hence C can detect t errors by the preceding discussion.

(2) If $2t + 1 \leq d$, it suffices (by the preceding discussion) to show that the t -balls about distinct code words are pairwise disjoint. But if $c \neq c'$ in C and $w \in S_t(c') \cap S_t(c)$, then the triangle inequality gives

$$d(c, c') \leq d(c, w) + d(w, c') \leq t + t = 2t < d$$

by the hypothesis, a contradiction. ■

Example 3. The following 7-code has minimum distance 3, so it can detect 2 errors and correct 1 error.

$$\{0000000, 0101010, 1010101, 1110000, 1011010, 0100101, 0001111, 1111111\}. \quad ■$$

⁴⁰If C can detect (correct) t or fewer errors, we say simply that C detects (corrects) t errors.

If c is any word in B^n , a word w satisfies $d(w, c) = r$ if and only if w and c differ in exactly r bits. Hence there are exactly $\binom{n}{r}$ such words w (where $\binom{n}{r}$ is the binomial coefficient), because there are $\binom{n}{r}$ ways to choose r bits of c to change. Therefore

$$|S_t(c)| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{t}.$$

This leads to a useful bound on the size of error-correcting codes.

Theorem 3. Hamming Bound. If an n -code C can correct t errors, then

$$|C| \leq \frac{2^n}{\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{t}}.$$

Proof. Write $N = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{t}$. The t -balls centered at distinct code words each contain N words, and there are $|C|$ of them. Hence they contain $N|C|$ distinct words (being pairwise disjoint). Hence $N|C| \leq 2^n$ because $|B^n| = 2^n$. This proves the theorem. \blacksquare

An n -code C is called **perfect** if there is equality in Theorem 3 or, equivalently, if every word in B^n lies in exactly one t -ball about a code word. Such codes exist. For example, if $n = 3$ and $t = 1$, then $\binom{3}{0} + \binom{3}{1} = 4$ and the Hamming bound is $2^3/4 = 2$. The 3-code $C = \{000, 111\}$ has minimum distance 3, so by Theorem 2 it can correct 1 error. Hence C is perfect. We present another example of a perfect code later.

Binary Linear Codes and Coset Decoding

Up to this point we have regarded any nonempty subset of B^n as an n -code. However, many important codes are subgroups. The group B^n has order 2^n so, by Lagrange's theorem, each subgroup has order 2^k for some $k = 0, 1, \dots, n$. Given integers k and n , with $1 \leq k \leq n$, an additive subgroup C of B^n of order 2^k is called an (n, k) -**binary linear code** (or simply an (n, k) -code). Note that we do not regard the trivial subgroup ($k = 0$) as a code.

Example 4. The code $\{00000, 11111\}$ in Example 1 is a $(5, 1)$ -code.

Example 5. The n -parity-check codes in Example 2 are $(n, n - 1)$ -codes, because the sum of two words of even parity also has even parity.

Example 6. $\{0000, 0101, 1010, 1111\}$ is a $(4, 2)$ -code. The following is a $(4, 3)$ -code: $\{0000, 0010, 0101, 0111, 1000, 1010, 1101, 1111\}$.

Many of the properties of the general n -codes take a simpler form for linear codes. The first part of the next theorem gives a much easier way to find the minimum distance of a linear code, the second and third parts strengthen Theorem 2, and the fourth part reformulates the Hamming bound.

Theorem 4. Let C be an (n, k) -code with minimum distance d .

- (1) $d = \min\{\text{wt } w \mid 0 \neq w \in C\}$.⁴¹
- (2) C can detect t errors if and only if $t + 1 \leq d$.

⁴¹Because of this the minimum distance of a linear code is sometimes called the *minimum weight* of the code.

- (3) C can correct t errors if and only if $2t + 1 \leq d$.
(4) If C can correct t errors, then $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{t} \leq 2^{n-k}$.

Proof. (1) Write $d' = \min\{\text{wt } w \mid 0 \neq w \in C\}$. If $0 \neq w \in C$, then $\text{wt } w \geq d$ by the definition of d . Hence $d' \geq d$. However, Theorem 1 gives $d(v, w) = \text{wt}(v - w)$ for all $v \neq w$ in C , so $d(v, w) \geq d'$ because $v - w \in C$ (C is a group). Hence $d \geq d'$.

(2) Assume that C can detect t errors. If $w \in C$, $w \neq 0$, the t -ball about w contains no other code word (see the discussion preceding Theorem 2). In particular, it does not contain the code word 0, so $t + 1 \leq d(w, 0) = \text{wt } w$. Hence $t + 1 \leq d$ by (1). The converse is part of Theorem 2.

(3) If C corrects t errors, the t -balls about code words are pairwise disjoint (see the discussion preceding Theorem 2). It suffices to show that $\text{wt } c \geq 2t + 1$ for all $c \in C$, $c \neq 0$, since then $d \geq 2t + 1$.

So assume, on the contrary, that $\text{wt } c \leq 2t$. We show that then $S_t(0) \cap S_t(c) \neq \emptyset$, a contradiction. Since $c \notin S_t(0)$, we have $\text{wt}(c) > t$, so c has more than t ones as bits. Form w by changing exactly t of these ones to zeros, and leaving the other bits of c as they were. Then $d(w, c) = t$, so $w \in S_t(c)$. But c has at most $2t$ ones as digits ($\text{wt}(c) \leq 2t$), so w will have at most t ones. Hence $\text{wt}(w) \leq t$; that is $d(w, 0) \leq t$; that is $w \in S_t(0)$. So $S_t(0) \cap S_t(c) \neq \emptyset$, as required.

- (4) Because $|C| = 2^k$, this assertion restates Theorem 3. ■

In practice, an (n, k) -code C contains a large number of words, so implementing nearest neighbor decoding by computing the distance between a received word and all 2^k code words is impractical at best. Fortunately, methods exist for reducing the amount of work required. One of these methods, called *coset decoding*, is based on the fact that the group B^n is partitioned into cosets by the subgroup C . In fact, there are $2^n/2^k = 2^{n-k}$ cosets $w + C$, where $w \in B^n$. The method depends on the following notion.

In each coset of C in B^n , choose a word e of minimum weight, called the **coset leader** for that coset. Note that there may be more than one candidate for coset leader. For example, if C is the code in Example 3 and $w = 0111000$, the coset

$$\begin{aligned} w + C = \{ &0111000, 0010010, 1101101, 1001000, \\ &1100010, 0011101, 0110111, 1000111 \} \end{aligned}$$

has two members of minimum weight 2.

After choosing the coset leaders, we can easily state the decoding procedure.

Coset Decoding Let C be an (n, k) -code. If a word $w \in B^n$ is received, and if e is any coset leader for $w + C$, decode w as $w - e$.

Theorem 5. Coset decoding is nearest neighbor decoding.

Proof. Let C be an (n, k) -code. If a word w is received and e is any coset leader in $w + C$, then $c = w - e$ is a code word in C (because e is in $w + C$). We must show that w is as close to c as any other element d of C . We have $w - d \in w + C = e + C$, so $\text{wt } e \leq \text{wt}(w - d)$ by the choice of e in C . Hence

$$d(w, c) = \text{wt}(w - c) = \text{wt } e \leq \text{wt}(w - d) = d(w, d),$$

which is what we wanted. ■

Example 7. Consider the $(6, 3)$ -code:

$$C = \{000000, 001110, 010101, 011011, 100011, 101101, 110110, 111000\}.$$

If $w = 101011$ and $v = 011100$ are received, decode them using coset decoding.

Solution. The cosets generated by w and v are

$$w + C = \{101011, 100101, 111110, 110000, 001000, 000110, 011101, 010011\}$$

$$v + C = \{011100, 010010, 001001, 000111, 111111, 110001, 101010, 100100\}$$

One of the coset leaders in $w + C$ is $e = 001000$, so w decodes as $w - e = 100011$. However, $v + C$ has three potential coset leaders: $f = 010010$, $g = 001001$, and $h = 100100$. These leaders decode v as 001110 , 010101 , and 111000 , respectively. Note that C has minimum distance 3, so it will correct one error by Theorem 4. Since w is one error away from 100011 (in C), the code corrects w . But $d(v, c) \geq 2$ for every word c in C , so the code does not correct v . Note that 001110 , 010101 , and 111000 are all the elements of C at distance 2 from v . \square

Given an (n, k) -code C for which $|C| = 2^k$ is not too large, we can carry out coset decoding by constructing a table (called a **standard array** for C), the rows of which are the various cosets $w + C$ of C in B^n . The coset $C = 0 + C$ is listed in the top row with 0 in column 1. (Note that 0 is the coset a for C .) In general, if e is any coset leader for $w + C$, then $w + C = e + C$, and we place the elements of this coset in a row of the table with e in column 1 and $e + c$ in the column headed by c for each $c \in C$. We then decode as follows: If we receive a word w , we locate it in the table (so $w = e + c$, where e is a coset leader) and decode it as the code word c at the head of its column. Here is an example.

Example 8. Construct the standard array for the $(4, 2)$ -code $C = \{0000, 0110, 1011, 1101\}$.

Solution. We obtain the rows

of this table as follows: The first row lists the elements of C in any order except that the coset leader 0 is in column 1; to obtain the next row, choose any element of B^4 not in C , say 1111, and construct the coset

$C = 0 + C$	0000	0110	1011	1101
$e_1 + C$	0100	0010	1111	1001
$e_2 + C$	1000	1110	0011	0101
$e_3 + C$	0001	0111	1010	1100

$$1111 + C = \{1111, 1001, 0100, 0010\}.$$

Next, we choose a coset leader, say, $e_1 = 0100$, (0010 would do as well), and obtain row 2 of the table by adding e_1 to the elements of row 1 in order. Thus, for example,

the word 1111 in column 3 is the sum of e_1 and the word 1011 (in C) at the head of column 3.

We complete the rest of the table in the same way. To form any row, we choose an element of B^4 not yet listed, find a coset leader in its coset, and list the coset as a row. The remaining coset leaders are $e_2 = 1000$ and $e_3 = 0001$ (each the unique word of minimum weight in its coset).

With the table complete, decoding is easy. For example, if we receive $w = 1010$, we decode w as $c = 1011$ because w is in column 3 of the table. \square

This method is impractical for large linear codes. For example, a $(40, 10)$ -code has $2^{30} > 10^9$ cosets, so finding the coset leaders is practically impossible. Hence large codes are constructed using more systematic methods.

Matrix Methods

One convenient way to obtain codes is by using matrix multiplication. Here we take the original messages to be the elements of B^k . We regard them as $1 \times k$ row matrices with entries from \mathbb{Z}_2 and encode by multiplying by a fixed binary matrix (entries from \mathbb{Z}_2). We use the usual rules for matrix multiplication, except that we do arithmetic modulo 2.

Example 9. The Hamming (7, 4)-code.⁴² We use the binary matrix

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

The message words are the elements of B^4 ; for example, $u = 1011$ is encoded as $uG = 1011001$ because of the matrix product

$$uG = [1 \ 0 \ 1 \ 1] \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} = [1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1].$$

In the chart below, the code words corresponding to all entries of B^4 appear on the right. Here each nonzero code word has weight at least 3, so the code can detect two errors and correct one error by Theorem 4.

⁴²This code was the first nontrivial example of an error correcting code given in the groundbreaking paper in which information theory was originated (Shannon, C.E., A mathematical theory of communication, *Bell Systems Technical Journal* 27 (1948), 623–656).

Message Word	Code Word
0000	0000000
0001	0001011
0010	0010101
0011	0011110
0100	0100110
0101	0101101
0110	0110011
0111	0111000
1000	1000111
1001	1001100
1010	1010010
1011	1011001
1100	1100001
1101	1101010
1110	1110100
1111	1111111

□

Observe that the first four columns of the matrix G in Example 9 form the 4×4 identity matrix I_4 . This ensures that the first four digits of each code word uG form the original message word. The general situation is described using the following terminology.

An (n, k) -code C is called a **systematic code** if each message word in B^k forms the first k digits of exactly one code word. A $k \times n$ matrix of the form⁴³

$$G = [I_k \ A]$$

is a **standard generator matrix** if I_k is the $k \times k$ identity matrix and A is a $k \times (n - k)$ binary matrix. Thus, the matrix G in Example 9 is a 4×7 standard generator matrix $G = [I_4 \ A]$ where

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

The code itself is given as $C = \{uG \mid u \in B^k\}$.

Theorem 6. Let G be a $k \times n$ standard generator matrix. Then

$$C = \{uG \mid u \in B^k\}$$

⁴³If A and B are $k \times m$ and $k \times n$ matrices, the notation $[A \ B]$ indicates the $k \times (m + n)$ matrix with A occupying the first m columns and B occupying the last n columns. The matrix $[A \ B]$ is said to be given in *block form*.

is a systematic (n, k) -code. Conversely, every systematic (n, k) -code is given in this way by a standard generator matrix G .

Proof. Define $\sigma : B^k \rightarrow B^n$ by $\sigma(u) = uG$ for all $u \in B^k$. Then σ is a group homomorphism because matrix multiplication satisfies $(u + v)G = uG + vG$. As σ is clearly onto C , this shows that C is a subgroup of B^n . In fact, σ is one-to-one. To see this, write $G = [I_k \ A]$, where A is $k \times (n - k)$. Then

$$\sigma(u) = u[I_k \ A] = [uI_k \ uA] = [u \ uA], \quad \text{for all } u \in B^k.$$

Hence σ is one-to-one because $\sigma(u) = \sigma(v)$ implies that $[u \ uA] = [v \ vA]$, when $u = v$. Thus B^k and C are isomorphic groups and, in particular, $|C| = |B^k| = 2^k$. Thus C is an (n, k) -code; it is systematic because $\sigma(u) = [u \ uA]$ for all $u \in B^k$. This proves the first part of Theorem 6; we leave the converse as Exercise 26. ■

Example 10. The $(6, 3)$ -code

$$C = \{000000, 001110, 010101, 011011, 100011, 101101, 110110, 111000\}$$

in Example 7 is systematic, and the reader can verify that it is generated by the standard generator matrix

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

That is, $C = \{uG \mid u \in B^3\}$. □

If C is a systematic (n, k) -code, we can easily write down a standard generator matrix for C . Because C is systematic, it contains a word c_i whose first k digits form row i of I_k . Let G be the $k \times n$ matrix whose rows are c_1, c_2, \dots, c_k in order:

$$G = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix}.$$

Then G is a standard generator matrix and $C = \{uG \mid u \in B^k\}$ (See Exercise 26). Incidentally, we say that C is **generated** by G when $C = \{uG \mid u \in B^k\}$. In this case C consists of 0 and all sums of (1 or more) of the generating words c_1, c_2, \dots, c_k . This is illustrated in Example 11.

Example 11. Both the codes $\{0000, 0101, 1010, 1111\}$ and

$$\{0000, 0010, 0101, 0111, 1000, 1010, 1101, 1111\}$$

in Example 6 are systematic with matrices $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$. □

On the other hand, the $(7, 3)$ -code in Example 3 is not systematic. Nonetheless, every (n, k) -code is *close* to being systematic in the sense that it contains k words with the following property: If F is the $k \times n$ matrix with these words as rows, F contains every column of the $k \times k$ identity matrix I_k (Exercise 27).

The use of a standard generator matrix is a convenient method of generating (n, k) -codes not only because retrieval is easy but also because a $k \times n$ matrix has only kn entries to store, whereas the code contains 2^k words of n entries each. Moreover, the process of encoding with a systematic code is simple: Multiply the message word by the generator matrix. Hence it is not surprising that matrix methods give a simple way to detect and correct errors.

To understand why, let C be a systematic binary (n, k) -code with standard generator matrix $G = [I_k \ A]$, where A is a $k \times (n - k)$ binary matrix. The **parity-check matrix**⁴⁴ for C is the $n \times (n - k)$ matrix given in block form by

$$H = \begin{bmatrix} A \\ I_{n-k} \end{bmatrix}.$$

If w is a word in B^n , the word wH in B^{n-k} is called the **syndrome** of w . Note that each of G and H completely determines the other, so either matrix determines the code C .

Example 12. The Hamming (7, 4)-code in Example 9 has the generator matrix

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \text{ where } A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Hence the parity-check matrix is $H = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. □

In Example 12, the reader can verify that $GH = 0$ is the zero matrix. This relation holds in general.

Lemma. *If G and H are the standard generator matrix and the parity-check matrix of a systematic (n, k) -code, then $GH = 0$.*

Proof. Write $G = [I_k \ A]$ so that $H = \begin{bmatrix} A \\ I_{n-k} \end{bmatrix}$. Then block multiplication gives

$$GH = [I_k \ A] \begin{bmatrix} A \\ I_{n-k} \end{bmatrix} = I_k A + A I_{n-k} = A + A = 0$$

where $A + A = 0$ because A is binary and $x + x = 0$ for all $x \in \mathbb{Z}_2$. ■

⁴⁴Systematic binary codes are often *defined* using the parity-check matrix. Then the transpose of H is referred to as the parity-check matrix.

Theorem 7. Orthogonality Theorem. Let C be a systematic (n, k) -code with parity-check matrix H .

- (1) $C = \{w \in B^n \mid wH = 0\}$.
- (2) Words w and v in B^n lie in the same C -coset if and only if $wH = vH$.

Proof. (1) Let $G = [I_k \ A]$ be the generator matrix for C , so $H = \begin{bmatrix} A \\ I_{n-k} \end{bmatrix}$. Define $\alpha : B^n \rightarrow B^{n-k}$ by $\alpha(w) = wH$ for all $w \in B^n$. Then α is a group homomorphism because $(w + v)H = wH + vH$, and (1) amounts to showing that $C = \ker \alpha$. We first verify that α is onto. If $v \in B^{n-k}$, let $w = [0 \ v] \in B^n$ be the word whose first k bits are zero and which ends with v . Then

$$\alpha(w) = wH = [0 \ v] \begin{bmatrix} A \\ I_{n-k} \end{bmatrix} = 0A + vI_{n-k} = 0 + v = v.$$

Hence α is onto, so $\text{im } \alpha = B^{n-k}$. Now the isomorphism theorem (Theorem 4 §2.10) gives $B^n / (\ker \alpha) \cong B^{n-k}$, so $|B^n| / |\ker \alpha| = |B^{n-k}|$. Therefore $|\ker \alpha| = 2^k$ and so $|\ker \alpha| = |C|$. Then to prove that $C = \ker \alpha$, it suffices to show that $C \subseteq \ker \alpha$. But if $c \in C$, then $c = uG$ for some $u \in B^k$ (Theorem 6), so $\alpha(c) = cH = uGH = u0 = 0$ by the lemma. Hence $C \subseteq \ker \alpha$.

- (2) For w and v in B^n , we have a chain of equivalences

$$w + C = v + C \Leftrightarrow w - v \in C \Leftrightarrow (w - v)H = 0 \Leftrightarrow wH = vH$$

where the first equivalence comes from Theorem 1 §2.6, the second is by (1), and the third is because $(w - v)H = wH - vH$. ■

The orthogonality theorem enables us to reformulate the coset decoding algorithm entirely in terms of the parity-check matrix.

Syndrome Decoding Let C be a systematic (n, k) -code with parity-check matrix H . If $w \in B^n$ is received, compute its syndrome wH and find a word $e \in B^n$ of minimal weight with the same syndrome (that is, $wH = eH$). Decode w as $c = w - e$.

The advantage of this method is that it requires knowing only the syndromes of the coset leaders (rather than the entire standard array), and sometimes the coset leaders can be discovered without finding the whole array.

Nearest neighbor decoding, as we have described it, is complete decoding in the sense that every received word is decoded. However, in many cases (especially where retransmission is easy) a better approach is to use a partial decoding procedure that corrects t errors and calls for retransmission when more than t errors are detected. We conclude by describing one such algorithm.

In this section, we have merely touched the surface of algebraic coding theory. For example, these results generalize with very little change if an (n, k) -code is defined to be a k -dimensional subspace of an n -dimensional vector space V over a finite field F (in our discussion, $V = B^n$ and $F = \mathbb{Z}_2$). Even more sophisticated

coding algorithms exist that use ring theory and field theory as well as group theory and linear algebra (see Section 6.7 for one such application).⁴⁵

Exercises 2.11

1. Find the Hamming weight of each word.
 - (a) 10110110
 - (b) 11010110
 - (c) 00101011011
 - (d) 010110101011
2. Find the Hamming distance between each pair of words.
 - (a) 101101 and 010101
 - (b) 10110101 and 01110111
 - (c) 1110111 and 0001000
 - (d) 10110111 and 01001011
3. Show that $d(v, w) = d(u + v, u + w)$ for all u, v , and w in B^n .
4. What is the maximum value of $d(v, w)$ when $v, w \in B^n$? Describe the pairs of words v and w in B^n with $d(v, w)$ as large as possible.
5. Let \bar{w} be the word obtained from $w \in B^n$ by changing every bit.
 - (a) Show that $\bar{v} + \bar{w} = v + w$ for all $v, w \in B^n$.
 - (b) Show that $d(v, w) + d(v, \bar{w}) = n$ for all $v, w \in B^n$.
6. Let C be the $(7, 3)$ -code in Example 3. Find the nearest neighbors to each of the following words in B^7 and so correct them (if possible).
 - (a) 0110101
 - (b) 0101110
 - (c) 1011001
 - (d) 1100110
7. How many errors can be detected or corrected by each of the following codes?
 - (a) $C = \{0000000, 0011110, 0100111, 0111001,$
 $1001011, 1010101, 1101100, 1110010\}$
 - (b) $C = \{0000000000, 001001111, 0101100111, 0111111000,$
 $1001110001, 1011101110, 1100010110, 1110001001\}$
8. Let c be a word in B^n and let $0 \leq t \leq n$. Show that $S_t(c) = \{v + c \mid v \in S_t(0)\}$.
9. (a) Show that the Hamming bound is equality if $t = 1$ in the $(7, 4)$ -Hamming code.
 (b) What is the maximum number of errors that an $(8, 3)$ -code can correct?
 (c) Is there a $(7, 2)$ -code of minimum distance 5?
10. (a) If a systematic $(n, 2)$ -code corrects one error, use the Hamming bound to show that $n \geq 5$ and find a $(5, 2)$ -code that corrects one error.
 (b) If a systematic $(n, 2)$ -code corrects two errors, use the Hamming bound to show that $n \geq 7$. Show that no $(7, 2)$ -code can correct two errors. Is there an $(8, 2)$ -code that corrects two errors? Justify your answer.
11. (a) If an $(n, 3)$ -code corrects two errors, show that $n \geq 9$.
 (b) Find a $(10, 3)$ -code that corrects two errors. It can be shown that there is no $(9, 3)$ -code that corrects two errors.
12. Given $r \geq 2$, write $n = 2^r - 1$ and $k = 2^r - r - 1$ so that $n - k = r$. Define H to be the $n \times r$ parity-check matrix consisting of all $n = 2^r - 1$ nonzero elements of B^r with I_r forming the last r rows. The corresponding (n, k) -code is called a **Hamming code**. (The $(7, 4)$ -Hamming code is the case $r = 3$). Show that every Hamming code corrects one error.

⁴⁵An introduction to the subject is given in Pless, V., *Introduction to the Theory of Error Correcting Codes*, New York: Wiley, 1982. A more thorough treatment (with an extensive bibliography) is that by MacWilliams, F.I., and Sloane, N.J.A., *The Theory of Error Correcting Codes*, Vols. I and II, New York: North Holland, 1977. Finally, a useful survey is contained in Chapter 4 of Lidl, R., and Pilz, G., *Applied Abstract Algebra*, New York: Springer-Verlag, 1984.

13. If a code word c is transmitted and w is received, show that coset decoding will correctly decode w if and only if $w - c$ is a coset leader in $w + C$.
14. Suppose that an (n, k) -code C has the property that each word $e \in B^n$, with $\text{wt } e \leq t$, is a coset leader in $e + C$. Show that C corrects t errors by using coset decoding.
15. (a) Show that no $(4, 2)$ -code can correct single errors.
 (b) Construct a $(5, 2)$ -code that can correct a single error.
16. (a) Show that no $(6, 3)$ -code can correct two errors.
 (b) Construct a $(6, 3)$ -code that can correct a single error.
 (c) Show that no $(7, 3)$ -code can correct two errors.
17. Given words v and w in B^n , define their product vw to be the word whose i th digit is the product $v_i w_i$ in \mathbb{Z}_2 , where v_i and w_i are the i th digits of v and w .
 (a) Show that $\text{wt}(v + w) + 2 \text{wt}(vw) = \text{wt } v + \text{wt } w$.
 (b) Deduce the triangle inequality: $\text{wt}(v + w) \leq \text{wt } v + \text{wt } w$. (See Theorem 1.)
 (c) Show that equality holds in (b) if and only if the i th bit of w is 0 whenever the i th bit of v is 1.
18. If $v, w \in B^n$, show that $\text{wt}(v + w) \geq \text{wt } v - \text{wt } w$ with equality if and only if the i th bit of v is 1 whenever the i th bit of w is 1. [Hint: Preceding exercise.]
19. If C is an (n, k) -code, $w \in B^n$ and $w \notin C$, show that $D = C \cup (w + C)$ is an $(n, k + 1)$ -code.
20. Write down the standard generator matrix G and the parity-check matrix H for each of the following systematic codes.
 (a) $C = \{00000, 11111\}$.
 (b) C = any systematic $(n, 1)$ -code.
 (c) The code in Exercise 7(a).
 (d) The code in Exercise 7(b).
21. List the codes generated by each standard generator matrix.
- | | |
|---|---|
| (a) $\begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ | (b) $\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$ |
| (c) $\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$ | (d) $\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}$ |
22. If C is the $(n, n - 1)$ -parity-check code (Example 2), show that C is systematic and describe the standard generating matrix G and the parity-check matrix H .
23. (Requires matrix algebra) Prove Theorem 7(a) without using the isomorphism theorem by writing each $w \in B^n$ such that $wH = 0$ as $w = [u \ v]$ where u consists of the first k bits of w , and v is the last $n - k$ bits of w .
24. (Requires matrix algebra) Let C and C' be (n, k) -codes, with standard generator matrices G and G' , and parity-check matrices H and H' , respectively.
 (a) Show that $C = C'$ if and only if $G = G'$.
 (b) Show that $C = C'$ if and only if $H = H'$.
25. Let C be an (n, k) -code.
 (a) Show that either each word in C has even weight or exactly half have even weight.
 (b) Show that either each word in C has n th bit 0 or exactly half have n th bit 0.
 (c) Generalize.
26. Show that every systematic (n, k) -code C is generated by a $k \times n$ standard generator matrix G ; that is, $C = \{uG \mid u \in B^k\}$. [Hint: Let c_1, c_2, \dots, c_k be the rows of I_k ; that is, c_i has the i th bit 1 and all other bits 0. If $[c_i \ c'_i]$ is the unique element of C with c_i as its first k bits, take $G = [I_k \ A]$, where the rows of A are c'_1, c'_2, \dots, c'_k .]

27. (Requires linear algebra) Show that every (n, k) -code C contains k words c_1, c_2, \dots, c_k

such that the $k \times n$ matrix $K = \begin{bmatrix} c_1 \\ \vdots \\ c_k \end{bmatrix}$ contains every column of the $k \times k$ identity

matrix I_k . [Hint: Regard C as a vector space over \mathbb{Z}_2 and let $\{b_1, \dots, b_k\}$ be a basis. If $B = \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix}$, carry B to reduced row-echelon form $B \rightarrow R$ and let c_i be row i of R .]

Chapter 3

Rings

Algebra is the intellectual instrument which has been created for rendering clear the quantitative aspect of the world.

—Alfred North Whitehead

Mathematics takes us still further from what is human into the region of absolute necessity, to which not only the actual world, but every possible world must conform.

—Bertrand Russell

Two of the earliest sources of the theory of rings lie in geometry and number theory. The study of surfaces determined by polynomial equations involved the addition and multiplication of polynomials in several variables. In addition, attempts to extend the prime factorization theorem for integers led to consideration of sets of complex numbers that were closed under addition and multiplication. Both cases involve a commutative multiplication. David Hilbert, who coined the term *ring*, and Richard Dedekind began the abstraction of these systems.

Earlier, in 1843, William Rowan Hamilton had introduced his quaternions. They are a noncommutative ring that contains the complex numbers, and he developed a calculus for them that he hoped would be useful in physics. At about the same time, Hermann Günther Grassmann was studying rings obtained by introducing a multiplication in what would today be called a finite dimensional vector space. The study of these “hypercomplex numbers” culminated in 1909 in the structure theorems of Joseph Henry MacLagan Wedderburn, which mark the beginning of noncommutative ring theory.

However, it was not until 1921 that Emmy Noether unified and simplified much of the work up to her time by applying “finiteness conditions” to rings. Her monumental work has, as B. L. van der Waerden observed, “had a profound effect on the

development of modern algebra.” In particular, in 1927 it motivated Emil Artin to prove a far reaching extension of Wedderburn’s theorem that influenced a whole generation of ring theorists. This result is presented in Chapter 11.

3.1 EXAMPLES AND BASIC PROPERTIES

The most commonly used algebraic systems are the sets \mathbb{Z} , \mathbb{R} , \mathbb{Q} , and \mathbb{C} of numbers, and they have *two* operations: they are closed under addition and multiplication. In this chapter we discuss such systems for which addition and multiplication satisfy many of the properties familiar from arithmetic.

A set R is called a **ring** if it has two binary operations, written as addition and multiplication, satisfying the following axioms for all a, b , and c in R .

- R1 $a + b = b + a$.
- R2 $a + (b + c) = (a + b) + c$.
- R3 An element 0 in R exists such that $0 + a = a$ for all a .
- R4 For each a in R an element $-a$ in R exists such that $a + (-a) = 0$.
- R5 $a(bc) = (ab)c$.
- R6 An element 1 in R exists such that $1 \cdot a = a = a \cdot 1$ for all a .
- R7 $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$.

And R is called a **commutative ring** if, in addition,

- R8 $ab = ba$ for all a and b in R .

The first four axioms assert that a ring R is an additive abelian group. The additive unity 0 in axiom R3 is called the **zero** of R , and the additive inverse $-a$ of a in axiom R4 is called the **negative** of the element a . Axioms R5 and R6 show that R is a multiplicative monoid, so the element 1 , called the **unity**⁴⁶ of R , is unique (Theorem 1 §2.1). Sometimes we write the zero and unity as 0_R and 1_R if the ring must be emphasized. The two identities in axiom R7 are called the **distributive laws**, and are the only axioms that connect addition and multiplication.

Several important examples satisfy all the axioms for a ring except possibly R6, the existence of a unity. We call them **general rings**⁴⁷. However, nearly all the examples mentioned in this book have a unity.

Example 1. Each of \mathbb{Z} , \mathbb{R} , \mathbb{Q} , and \mathbb{C} is a commutative ring; \mathbb{Z}_n is a commutative ring for each $n \geq 2$ by Theorem 4 §1.3.

Example 2. The set $M_2(\mathbb{R})$ of all 2×2 matrices over \mathbb{R} is a ring using matrix addition and multiplication (see Appendix B). Note that $M_2(\mathbb{R})$ is noncommutative. Indeed, if $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ then $AB = B$ but $BA = 0$.

Example 3. The set $\mathbb{R}[x]$ of all polynomials with coefficients in \mathbb{R} is a ring with the usual addition and multiplication. We discuss such rings in detail in Chapter 4.

⁴⁶Other commonly used terms are **unit element** and **identity**.

⁴⁷Many authors use the term *ring* even if there is no unity and employ the term *ring with unity* when a unity exists. Our terminology is gaining acceptance because many examples of interest do indeed have a unity.

Example 4. If X is a nonempty set, let $F(X, \mathbb{R})$ be the set of all real valued functions $f: X \rightarrow \mathbb{R}$. Then $F(X, \mathbb{R})$ is a commutative ring using **pointwise addition** and multiplication: If f and g are in $F(X, \mathbb{R})$, we define

$$\begin{aligned} f + g : X \rightarrow \mathbb{R} &\quad \text{by} \quad (f + g)(x) = f(x) + g(x), \quad \text{for all } x \in X, \\ f \cdot g : X \rightarrow \mathbb{R} &\quad \text{by} \quad (f \cdot g)(x) = f(x)g(x), \quad \text{for all } x \in X. \end{aligned}$$

The zero of $F(X, \mathbb{R})$ is the constant function $\theta : X \rightarrow \mathbb{R}$ given by $\theta(x) = 0$ for all $x \in X$; the negative of $f \in F(X, \mathbb{R})$ is $-f : X \rightarrow \mathbb{R}$, defined by $(-f)(x) = -f(x)$ for all $x \in X$; and the unity is the constant function $1 : X \rightarrow \mathbb{R}$ defined by $1(x) = 1$ for all $x \in X$. We leave the routine verification of the other axioms the reader. \square

Example 5. If R_1, R_2, \dots, R_n are rings, we define **componentwise** operations on the cartesian product $R_1 \times R_2 \times \dots \times R_n$ as follows:

$$\begin{aligned} (r_1, r_2, \dots, r_n) + (s_1, s_2, \dots, s_n) &= (r_1 + s_1, r_2 + s_2, \dots, r_n + s_n) \\ (r_1, r_2, \dots, r_n) \cdot (s_1, s_2, \dots, s_n) &= (r_1 s_1, r_2 s_2, \dots, r_n s_n) \end{aligned}$$

Then $R_1 \times R_2 \times \dots \times R_n$ is a ring, the **direct product** of the rings R_1, R_2, \dots, R_n , and it is commutative if and only if each R_i is commutative. The additive group is just the direct product of the (additive) groups R_i , the unity is $(1, 1, \dots, 1)$, and the zero is $(0, 0, \dots, 0)$. \square

In the ring \mathbb{R} the property $0 \cdot r = 0$ for all r is important and highlights the unique multiplicative role played by 0. In fact, this property holds for every ring R . Because it involves the multiplication of R , and because 0 is the *additive* identity for R , it is not surprising that it is a consequence of the distributive laws.

Theorem 1. If 0 is the zero of a ring R , then $0r = 0 = r0$ for every $r \in R$.

Proof. Given $r \in R$, compute: $0r + 0r = (0 + 0)r = 0r = 0r + 0$. Hence, $0r = 0$ follows by adding $-0r$ to both sides (in the additive group R). Similarly, $r0 = 0$. \blacksquare

If it happens that $1 = 0$ in a ring R then, for any $r \in R$, $r = r \cdot 1 = r \cdot 0 = 0$ by Theorem 1. Hence, $R = \{0\}$ is the zero ring in Example 6.

Example 6. The set $R = \{0\}$ is a ring where $0 + 0 = 0$ and $0 \cdot 0 = 0$. It is called the **zero ring** and denoted $R = 0$.

Theorem 1 allows us to define matrix rings over an arbitrary ring.

Example 7. If $n \geq 1$, an $n \times n$ **matrix** over a ring R is an $n \times n$ array

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

where each a_{ij} is an element of R called the (i,j) -entry of A .

The set of all $n \times n$ matrices over R is denoted $M_n(R)$.

As for numerical matrices, we define equality, addition, and multiplication in $M_n(R)$ as follows: If $A = [a_{ij}]$ and $B = [b_{ij}]$ are in $M_n(R)$, then

- (1) $A = B$ if and only if $a_{ij} = b_{ij}$ for all i and j .
- (2) $A + B = [a_{ij} + b_{ij}]$.
- (3) $AB = [c_{ij}]$ where $c_{ij} = \sum_{k=1}^n a_{ik}b_{kj}$ for all i and j .

These are called **matrix operations**. It is routine to verify that $M_n(R)$ is an additive abelian group, the zero being the **zero matrix** $0 = [0]$ with each entry zero, and the **negative** of $A = [a_{ij}]$ is $-A = [-a_{ij}]$. The associative and distributive laws (axioms R5 and R7) follow from the corresponding properties of R (see Appendix B). The unity of $M_n(R)$ is the $n \times n$ **identity matrix**

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

with ones on the **main diagonal** (upper left to lower right) and zeros elsewhere. Hence $M_n(R)$ is a ring, called the **$n \times n$ matrix ring** over R . Note that, if $n \geq 2$, then $M_n(R)$ is noncommutative for every ring $R \neq 0$ (see Example 2). Thus, for example, $M_2(\mathbb{Z}_2)$ is a noncommutative ring with 16 elements. \square

Because a ring R is an *additive* abelian group, the laws of exponents take on a different form: If $n \in \mathbb{Z}$ and $a \in R$, we write the n^{th} “power” a^n of a as the n^{th} “multiple” na of a in an additive group. The following expressions translate other facts about exponents to additive notation.

$$\begin{array}{ll} 0a = 0 & a^0 = 1 \\ 1a = a & a^1 = a \\ (-1)a = -a & a^{(-1)} = a^{-1} \\ (n+m)a = na + ma & a^{n+m} = a^n a^m \\ n(a+b) = na + nb & (ab)^n = a^n b^n \text{ (if } ab = ba\text{)} \\ n(ma) = (nm)a & (a^m)^n = a^{mn} \end{array}$$

We use these formulas frequently without further comment.

If 1_R denotes the unity of R , we also write $1_R + 1_R = 2$, $1_R + 1_R + 1_R = 3$, and so on. More generally, we write

$$k \cdot 1_R = k, \quad \text{for all integers } k$$

when no confusion can result. This notation is consistent with our convention of writing $\mathbb{Z}_n = \{0, 1, 2, \dots, n-1\}$. Of course we are interested in how this “multiplication” by integers relates to the multiplication in R . As in Theorem 1, this depends on the distributive laws.

Theorem 2. *Let r and s be arbitrary elements of a ring R .*

- (1) $(-r)s = r(-s) = -(rs)$.
- (2) $(-r)(-s) = rs$.
- (3) $(mr)(ns) = mn(rs)$ for all integers m and n .

Proof. Theorem 1 gives $(-r)s + rs = (-r + r)s = 0s = 0 = -(rs) + rs$. Hence, $(-r)s = -(rs)$ by cancellation. Similarly, $r(-s) = -(rs)$, proving (1). Now (1) gives $(-r)(-s) = r[-(-s)] = rs$, proving (2). Turning to (3), we begin with

$$r(ns) = n(rs), \quad \text{for all } n \in \mathbb{Z}. \quad (*)$$

This holds for $n = 0$ by Theorem 1. If it holds for some $n \geq 0$, then

$$r[(n+1)s] = r(ns + s) = r(ns) + rs = n(rs) + 1(rs) = (n+1)rs.$$

Hence, (*) holds for all $n \geq 0$ by induction. If $n < 0$, write $n = -m$, $m > 0$. Then

$$r(ns) = r[-(ms)] = -[r(ms)] = -[m(rs)] = n(rs),$$

which proves (*). We leave it to the reader to show that $(mr)s = m(rs)$ holds for all $m \in \mathbb{Z}$, and that this equation and (*) imply (3). ■

If r and s are elements of a ring R , their **difference** $r - s$ is defined by

$$r - s = r + (-s).$$

Thus, the equation $x + s = r$ in R has the unique solution $x = r - s$. As for numbers, we say that $r - s$ is the result of **subtracting** s from r . Theorem 2 then gives the following extensions of the distributive laws:

$$a(b - c) = ab - ac \quad \text{and} \quad (b - c)a = ba - ca.$$

These expressions allow us to use the familiar properties of subtraction in any ring.

If R is any ring, we define the **characteristic** of R , denoted $\text{char } R$, in terms of the order of 1_R in the additive group $(R, +)$:

$$\begin{aligned} \text{char } R = n &\quad \text{if } o(1_R) = n \text{ in the additive group } (R, +), \\ \text{char } R = 0 &\quad \text{if } o(1_R) = \infty \text{ in the additive group } (R, +). \end{aligned}$$

If k is an integer, we write $kR = 0$ to mean that $kr = 0$ for each $r \in R$. By Theorem 2, this happens if and only if $k1_R = 0$ (verify), so we obtain

Theorem 3. *If R is a ring and $\text{char } R = n$ then*

- (1) *If $\text{char } R = n > 0$, then $kR = 0$ if and only if n divides k .*
- (2) *If $\text{char } R = 0$, then $kR = 0$ if and only if $k = 0$.*

Example 8. Each of \mathbb{Z} , \mathbb{R} , \mathbb{Q} , and \mathbb{C} has characteristic 0. Given $n \geq 2$, the ring \mathbb{Z}_n has characteristic n .

The binomial theorem for real variables (Example 6 §1.1) has a wide ranging generalization that will be needed later.

Theorem 4. Binomial Theorem. *Let a and b be elements in a ring R which commute, that is $ab = ba$. Then, for each $n \geq 0$*

$$(a + b)^n = \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + \binom{n}{n}b^n,$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ denotes the binomial coefficient (see Section 1.1).

Proof. It holds if $n = 0$ because $r^0 = 1$ for all $r \in R$, and it holds for $n = 1$ because $\binom{n}{0} = 1 = \binom{n}{n}$ for each $n \geq 0$. If it holds for some $n \geq 1$, compute

$$\begin{aligned}(a+b)^{n+1} &= (a+b)(a+b)^n \\&= (a+b) \left[\binom{n}{0} a^n + \binom{n}{1} a^{n-1} b + \cdots + \binom{n}{n-1} a b^{n-1} + \binom{n}{n} b^n \right] \\&= a^{n+1} + [\binom{n}{0} + \binom{n}{1}] a^n b + \cdots + [\binom{n}{n-1} + \binom{n}{n}] a b^n + b^{n+1} \\&= \binom{n+1}{0} a^{n+1} + \binom{n+1}{1} a^n b + \cdots + \binom{n+1}{n} a b^n + \binom{n+1}{n+1} b^{n+1}\end{aligned}$$

using the Pascal identity $\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}$ for $1 \leq k \leq n$ (Exercise 13 §1.1). This completes the induction, and so proves the binomial theorem. ■

Thus, for example, taking $a = b = 1$ and writing $1 + 1 = 2$, we obtain

$$2^n = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n-1} + \binom{n}{n} \text{ in any ring.}$$

Subrings

If R is a ring, a subset S is called a **subring** of R if it is itself a ring with the same operations (including the same unity)⁴⁸ as R . Thus, a subring of R is an additive subgroup of R that contains the unity of R and is closed under multiplication. The subgroup test (Theorem 1 §2.3) then gives

Theorem 5. Subring Test. A subset S of a ring R is a subring if and only if the following conditions are satisfied.

- (1) $0 \in S$ and $1 \in S$.
- (2) If $s \in S$ and $t \in S$, then $s + t$, st , and $-s$ are all in S .

As S is nonempty by (1), note that (2) is equivalent to the following condition: If $s \in S$ and $t \in S$, then $st \in S$ and $s - t \in S$.

Example 9. If $i^2 = -1$ in \mathbb{C} , write $\mathbb{Z}(i) = \{n + mi \in \mathbb{C} \mid m, n \in \mathbb{Z}\}$. Then $\mathbb{Z}(i)$ is a subring of \mathbb{C} by the subring test, called the ring of **gaussian integers**.

Example 10. If R is any ring, let

$$T_2(R) = \left\{ \begin{bmatrix} R & R \\ 0 & R \end{bmatrix} \mid a, b, c \text{ in } R. \right\}$$

Show that $T_2(R)$ is a subring of $M_2(R)$. This is called the ring of **upper triangular matrices** over R .

Solution. Clearly, the 2×2 zero matrix and the 2×2 identity matrix are in $T_2(R)$. Given $A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$ and $B = \begin{bmatrix} p & q \\ 0 & r \end{bmatrix}$ in $T_2(R)$, it is enough, by the subring test, to

⁴⁸The term *subring* is sometimes used for a general ring contained in R (possibly with a unity different from that of R). However, we insist that S and R have the same unity.

observe that each of the following matrices is in $T_2(R)$:

$$-A = \begin{bmatrix} -a & -b \\ 0 & -c \end{bmatrix}, \quad A + B = \begin{bmatrix} a+p & b+q \\ 0 & c+r \end{bmatrix}, \quad AB = \begin{bmatrix} ap & aq+br \\ 0 & cr \end{bmatrix}. \quad \square$$

The ring of upper triangular matrices will be referred to again. In general, subrings of $M_2(R)$ are a fertile source of interesting examples of rings.

Example 11. The set of continuous functions $\mathbb{R} \rightarrow \mathbb{R}$ is an important subring of $F(\mathbb{R}, \mathbb{R})$ (see Example 4). Closure under addition, multiplication, and negation are theorems of calculus. The differentiable functions are also a subring of $F(\mathbb{R}, \mathbb{R})$.

Example 12. If R is a ring, the **center** $Z(R)$ of R is a subring of R , where

$$Z(R) = \{z \in R \mid zr = rz \text{ for all } r \in R\}.$$

Verification is left to the reader. Elements in $Z(R)$ are said to be **central** in R . \square

An element e in a ring R is called an **idempotent** if $e^2 = e$. Examples of idempotents include 0 and 1 in any ring R , $(1, 0)$ and $(0, 1)$ in $\mathbb{R} \times \mathbb{R}$, $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ in $M_2(\mathbb{R})$, and 3 and 4 in \mathbb{Z}_6 . If e is any idempotent, so also is $(1 - e)$, and $e(1 - e) = 0 = (1 - e)e$. If R is a ring, the idempotents in R provide an important class of rings S contained in R , (using the operations of R) that may not be subrings (that is, they may have a different unity).

Theorem 6. If $e = e^2$ in a ring R , write $eRe = \{ere \mid r \in R\}$. Then eRe is a ring with unity e , and $eRe = \{a \in R \mid ea = a = ae\}$.

Proof. We have $0 = e0e$, $ere + ese = e(r + s)e$, and $-ere = e(-r)e$. Hence, eRe is an additive subgroup of R that is closed under multiplication. Finally, the fact that $e^2 = e$ means that $e = eee \in eRe$ and $e(ere) = ere = (ere)e$. Thus e is the unity of eRe , and $eRe \subseteq \{a \in R \mid ae = a = ea\}$. The reverse inclusion is easy. \blacksquare

The rings eRe , $e^2 = e \in R$, are called **corners** of the ring R . The next example explains the name.

Example 13. If $e = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \in R = M_2(\mathbb{R})$, then $e^2 = e$ and $eRe = \left\{ \begin{bmatrix} r & 0 \\ 0 & 0 \end{bmatrix} \mid r \in \mathbb{R} \right\}$.

Units and Division Rings

If R is any ring, an element u in R is called a **unit** if u has a multiplicative inverse in R (denoted u^{-1}). Thus $u(u^{-1}) = 1 = (u^{-1})u$. The units in a ring R can be “canceled”. More precisely, if u is a unit in R and $r, s \in R$ then

$$ur = us \text{ implies } r = s \quad \text{and} \quad ru = su \text{ implies } r = s.$$

by left and right multiplication by u^{-1} .

The set of all units in R , denoted R^* , is a multiplicative group (Theorem 1 §2.2) called the **group of units** of the ring R . This terminology is consistent with the

notation M^* for the units in any multiplicative monoid. For example, $\mathbb{Z}^* = \{1, -1\}$, and $\mathbb{Z}_n^* = \{\bar{k} \mid \gcd(k, n) = 1\}$ by Theorem 5 §1.3.

Example 14. Show that $\mathbb{Z}(i)^* = \{1, -1, i, -i\}$, where $\mathbb{Z}(i) = \{a + bi \mid a, b \in \mathbb{Z}\}$ is the ring of gaussian integers (Example 9).

Solution. Let $u = a + bi$ be a unit in $\mathbb{Z}(i)$. Because $uu^{-1} = 1$, taking absolute values gives $|u||u^{-1}| = 1$. Hence, $|u| = 1$ so $a^2 + b^2 = |u|^2 = 1$. As a and b are also integers, the only solutions are $a = \pm 1$, $b = 0$, or $a = 0$, $b = \pm 1$, so $u = 1, -1, i$, or $-i$. \square

Example 15. Let R be a commutative ring. If A is an $n \times n$ matrix in $M_n(R)$, the determinant $\det A$ of A is defined exactly as in $M_n(\mathbb{R})$, and satisfies

$$\det A \det B = \det AB \quad \text{for all } A, B \in M_n(R).$$

Then the usual linear algebra argument (when $R = \mathbb{R}$) gives

$$M_n(R)^* = \{A \in M_n(R) \mid \det A \in R^*\}.$$

In addition, the **adjugate** $\text{adj } A$ is defined, again exactly as in $M_n(\mathbb{R})$, and we have

$$\text{If } A \in M_n(R) \text{ and } \det A \in R^*, \text{ then } A^{-1} = (\det A)^{-1} \text{adj } A.$$

We emphasize that R *must* be commutative. This is discussed in Appendix B.

For example, if $n = 2$, $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and $\det A = ad - bc$ is a unit in R , then $A^{-1} = (\det A)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ as can be readily verified directly. \square

If R is not the zero ring, the zero element 0 is *never* a unit since otherwise $1 = 0^{-1}0 = 0$ by Theorem 1. Hence, every unit in R must be nonzero, and the rings where the converse holds are very important. A ring $R \neq 0$ is called a **division ring**⁴⁹ if every nonzero element of R is a unit in R ; that is, if $R^* = R \setminus \{0\}$. A commutative division ring is called a **field**.

Example 16. \mathbb{Q} , \mathbb{R} , and \mathbb{C} are fields. If p is a prime, \mathbb{Z}_p is a field by Theorem 7 §1.3. Note that \mathbb{Z} is *not* a field.

For now, we have no other examples of fields and no examples at all of noncommutative division rings. We have more to say about them in the next section.

An element a in a ring R is said to be **nilpotent** (and is called a **nilpotent element**) if $a^k = 0$ for some $k \geq 1$. Clearly, 0 is a nilpotent in every ring. Other examples of nilpotents include 2 and 4 in \mathbb{Z}_8 , $\begin{bmatrix} 0 & r \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ r & 0 \end{bmatrix}$ in $M_2(R)$, and $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ in $M_2(\mathbb{Z}_2)$. The observation in Example 17 is often useful.

⁴⁹Also called a **skew field**.

Example 17. If a is a nilpotent in R , show that $1 - a$ and $1 + a$ are units.

Solution. If $a^n = 0$ where $n \geq 1$, then $a^k = 0$ for all $k \geq n$ by Theorem 1. Hence $u = 1 + a + a^2 + a^3 + \dots$ is a finite sum, and so is an element of R . Then $(1 - a)u = 1 = u(1 - a)$, as the reader can verify. Hence $1 - a$ is a unit. As $(-a)^n = 0$, $1 + a$ is a unit too. \square

In elementary algebra it is proved that, if $x \in \mathbb{R}$, and $|x| < 1$, the geometric series $1 + x + x^2 + x^3 + \dots$ converges for any real number x with $|x| < 1$ and equals $(1 - x)^{-1}$ in this case. In the solution to Example 17 we recognize that $1 + a + a^2 + \dots$ makes sense in *any* ring R when a is a nilpotent, which then provides a formula for $(1 - a)^{-1}$. This argument makes sense for other power series provided the coefficients are in the ring R , say $e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots$, assuming $2, 3, \dots$ are all units.

Example 18. If a ring R has no nonzero nilpotent elements, show that all idempotents in R are central.

Solution. If $e^2 = e \in R$ and $r \in R$, write $a = er(1 - e)$. Then $a^2 = 0$ because $(1 - e)e = 0$, so $a = 0$ by hypothesis. It follows that $er = ere$. Similarly $re = ere$, so $er = ere = er$. See Exercises 22 and 23 for other such properties. \square

Ring Isomorphisms

The concept of isomorphic rings is analogous to the corresponding notion for groups, and it is equally important. Two rings R and S are called **isomorphic** (written $R \cong S$) if there is a mapping $\sigma : R \rightarrow S$ that satisfies the following conditions.

- (1) σ is a bijection.
- (2) $\sigma(r + s) = \sigma(r) + \sigma(s)$ for all r and s in R . σ preserves sums
- (3) $\sigma(rs) = \sigma(r) \cdot \sigma(s)$ for all r and s in R . σ preserves products

Such a map σ is called a **ring isomorphism**. An isomorphism $R \rightarrow R$ is called an **automorphism** of R .

Conditions (1) and (2) in the definition show that a ring isomorphism $\sigma : R \rightarrow S$ is also an isomorphism of additive groups, so it also preserves zero, negatives, and \mathbb{Z} -multiples (Theorem 1 §2.5). That is, for $r \in R$:

$$\sigma(0) = 0, \quad \sigma(-r) = -\sigma(r), \quad \text{and} \quad \sigma(kr) = k\sigma(r) \text{ for all } k \in \mathbb{Z}.$$

In addition, σ preserves the unity; that is

$$\sigma(1_R) = 1_S.$$

To see why, write $\sigma(1_R) = e$, and let $s \in S$. Since σ is onto, we have $s = \sigma(r)$ for some $r \in R$, so

$$se = \sigma(r) \cdot \sigma(1_R) = \sigma(r \cdot 1_R) = \sigma(r) = s.$$

Similarly $es = s$, so $e = 1_S$ is the unity of S . Moreover, conditions (2) and (3) above show that σ preserves the addition and multiplication tables and hence, as for groups, isomorphic rings R and S are the same except for the notations used.

Example 19. If R is a ring and $u \in R^*$ is a unit in R , define

$$\sigma_u : R \rightarrow R \quad \text{by} \quad \sigma_u(r) = uru^{-1} \text{ for all } r \in R.$$

Then σ_u is an automorphism of the ring R called the **inner automorphism** determined by u . Indeed, σ_u preserves addition and multiplication because

$$u(r+s)u^{-1} = uru^{-1} + usu^{-1} \quad \text{and} \quad u(rs)u^{-1} = uru^{-1}usu^{-1}$$

for all $r, s \in R$. The proof that σ_u is one-to-one and onto is left to the reader. \square

These inner automorphisms are important. For example, two matrices A and B in $M_n(\mathbb{R})$ are called **similar** if $PAP^{-1} = B$ for an invertible matrix P , that is $B = \sigma_P(A)$. This is a fundamental concept in linear algebra.

Example 20. Show that $R = \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{R} \right\}$ is isomorphic to \mathbb{C} .

Solution. The reader can verify that R is a subring of $M_2(\mathbb{R})$. Define $\sigma : R \rightarrow \mathbb{C}$ by $\sigma \begin{bmatrix} a & -b \\ b & a \end{bmatrix} = a + bi$. Then σ is clearly onto; it is one-to-one because $a + bi = a' + b'i$ in \mathbb{C} means that $a = a'$ and $b = b'$; and it preserves addition (verify). Finally,

$$\begin{aligned} \sigma \left\{ \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} a' & -b' \\ b' & a' \end{bmatrix} \right\} &= \sigma \begin{bmatrix} aa' - bb' & -ab' - ba' \\ ba' + ab' & aa' - bb' \end{bmatrix} \\ &= (aa' - bb') + (ab' + ba')i \\ &= (a + bi)(a' + b'i) \\ &= \sigma \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \sigma \begin{bmatrix} a' & -b' \\ b' & a' \end{bmatrix} \end{aligned}$$

shows that σ preserves multiplication. Hence, σ is a ring isomorphism. \square

Example 21. Show that the rings $R = \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in \mathbb{Z}_2 \right\}$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are *not* isomorphic as rings, even though they are isomorphic as additive groups.

Solution. The element $r = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ in R satisfies $r^2 = 0$. Suppose $\sigma : R \rightarrow \mathbb{Z}_2 \times \mathbb{Z}_2$ is a ring isomorphism. If $s = \sigma(r)$, then $s^2 = \sigma(r)^2 = \sigma(0) = 0$. But $s^2 = 0$ in $\mathbb{Z}_2 \times \mathbb{Z}_2$ implies that $s = 0$, giving $r = 0$ because σ is one-to-one. This is a contradiction, so no such isomorphism σ can exist. However, the map $\begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mapsto (a, b)$ is an isomorphism of additive groups, as the reader can verify. \square

One of the consequences of Lagrange's theorem is that every group of prime order must be cyclic. We conclude with the analogue for rings.

Theorem 7. If $R \neq 0$ is a ring and $|R| = p$ is a prime, then $R \cong \mathbb{Z}_p$ is a field.

Proof. Define $\theta : \mathbb{Z}_p \rightarrow R$ by $\theta(\bar{k}) = k1_R$. This is well defined; in fact

$$\bar{k} = \bar{m} \text{ in } \mathbb{Z}_p \Leftrightarrow p|(k - m) \Leftrightarrow (k - m)1_R = 0 \Leftrightarrow k1_R = m1_R \text{ in } R,$$

so θ is well defined and one-to-one. Finally, 1_R is a generator of $(R, +)$ by Lagrange's theorem (p is a prime), which shows that θ is onto. Hence, θ is an isomorphism. \blacksquare

Exercises 3.1

Throughout these exercises R denotes a ring unless otherwise specified.

1. In each case explain why R is not a ring.
 - (a) $R = \{0, 1, 2, 3, \dots\}$, operations of \mathbb{Z}
 - (b) $R = 2\mathbb{Z}$
 - (c) $R =$ the set of all mappings $f : \mathbb{R} \rightarrow \mathbb{R}$; addition as in Example 4 but using composition as the multiplication
2. If R is a ring, define the **opposite ring** R^{op} to be the set R with the same addition but with multiplication $r \cdot s = sr$. Show that R^{op} is a ring.
3. In each case show that S is a subring of R .
 - (a) $S = \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \mid a, b, c, d \in \mathbb{R}, a + c = b + d \right\}, R = M_2(\mathbb{R})$
 - (b) $S = \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in \mathbb{R} \right\}, R = M_2(\mathbb{R})$
 - (c) $S = \left\{ \begin{bmatrix} a & 0 & b \\ 0 & c & d \\ 0 & 0 & a \end{bmatrix} \mid a, b, c, d \in \mathbb{R} \right\}, R = M_3(\mathbb{R})$
 - (d) $S = \left\{ \begin{bmatrix} a & 2b \\ b & a \end{bmatrix} \mid a, b \in \mathbb{R} \right\}, R = M_2(\mathbb{R})$
4. If S and T are subrings of R , show that $S \cap T$ is a subring. Is this the case for $S + T = \{s + t \mid s \in S, t \in T\}$?
5. If X is a nonempty subset of R , show that $C(X) = \{c \in R \mid cx = xc \text{ for all } x \in X\}$ is a subring of R (called the **centralizer** of X in R).
6. (a) If $ab = 0$ in a division ring R , show that $a = 0$ or $b = 0$.
 - (b) If $a^2 = b^2$ in a field, show that $a = b$ or $a = -b$.
7. Compute $Z[M_2(R)]$ for any ring R .
8. (a) Show that $(a+b)(a-b) = a^2 - b^2$ in a ring R if and only if $ab = ba$.
 - (b) Show that $(a+b)^2 = a^2 + 2ab + b^2$ in a ring R if and only if $ab = ba$.
9. Show that $a + b = b + a$ follows from the other ring axioms, where we assume that both $0 + a = a$ and $a + 0 = a$ hold for all a in R .
10. (a) If $ab + ba = 1$ and $a^3 = a$ in a ring, show that $a^2 = 1$.
 - (b) If $ab = a$ and $ba = b$ in a ring, show that $a^2 = a$ and $b^2 = b$.
11. Show that 0 is the only nilpotent in R if and only if $a^2 = 0$ implies $a = 0$.
12. If $a \neq b$ in R satisfy $a^3 = b^3$ and $a^2b = b^2a$, show that $a^2 + b^2$ is not a unit.
13. If u, v , and $u + v$ are all units in a ring R , show that $u^{-1} + v^{-1}$ is also a unit and give a formula for $(u^{-1} + v^{-1})^{-1}$ in terms of u, v , and $(u + v)^{-1}$. [Hint: Compute $u(u^{-1} + v^{-1})v$.]
14. Given r and s in a ring R , show that $1 + rs$ is a unit if and only if $1 + sr$ is a unit. [Hint: $s(1 + rs) = (1 + sr)s$.]
15. Show that the following conditions are equivalent for a general ring R .
 - (1) R has a unity.
 - (2) R has a right unity ($re = r$ for all r) and $Ra = 0$, $a \in R$, implies that $a = 0$.
 - (3) R has a unique right unity.

16. If 1_R denotes the unity of a ring R , write $\mathbb{Z}1_R = \{k1_R \mid k \in \mathbb{Z}\}$.
- Show that $\mathbb{Z}1_R$ is a subring of R contained in $\mathbb{Z}(R)$.
 - If $\text{char } R = n$, show that $\mathbb{Z}1_R \cong \mathbb{Z}_n$.
 - If $\text{char } R = 0$, show that $\mathbb{Z}1_R \cong \mathbb{Z}$.
17. Describe the rings of characteristic 1.
18. In each case, find the characteristic of the ring.
- $\mathbb{Z}_n \times \mathbb{Z}_m$
 - $M_2(\mathbb{Z}_n)$
 - $\mathbb{Z} \times \mathbb{Z}_n$
19. If u is a unit in R and $\text{char } R < \infty$, show that $\text{char } R = o(u)$ in $(R, +)$.
20. If $ua = au$, where u is a unit and a is a nilpotent, show that $u + a$ is a unit.
21. (a) If $e^2 = e$ in R , show that $1 - 2e$ is a unit, indeed self-inverse.
(b) If $2 \in R^*$, show that $\sigma : \{e \mid e^2 = e\} \rightarrow \{u \mid u^2 = 1\}$ is a bijection if $\sigma(e) = 1 - 2e$.
22. (a) If $e^2 = e$, show that $(1 - e)re$ and $er(1 - e)$ are nilpotents for all $r \in R$.
(b) If $e^2 = e$, show that $e + (1 - e)re$ and $e + er(1 - e)$ are idempotents for all $r \in R$.
(c) If $e^2 = e$, show that $1 + (1 - e)re$ and $1 + er(1 - e)$ are units for all $r \in R$.
23. Show that the following are equivalent for an idempotent $e^2 = e \in R$.
- e is central.
 - $ef = fe$ whenever $f^2 = f$.
 - $ea = ae$ for every nilpotent a
 - $eu = ue$ for every unit u
- [Hint: Exercise 22.]
24. Consider the following conditions on R : (1) every unit is central, (2) every nilpotent is central, and (3) every idempotent is central. Show that (1) \Rightarrow (2) \Rightarrow (3). [Hint: Exercise 22.]
25. If $r^3 = r$ for all $r \in R$, show that R is commutative. [Hint: Use Example 18 and Exercise 23 to show that a^2 central for all a .] Remark: In fact, Jacobson's theorem asserts that R is commutative if, for each $r \in R$, some $n \geq 2$ exists with $r^n = r$.
26. In each case show that $ab = 1$ in R implies that $ba = 1$.
- R is finite. [Hint: If $R = \{r_1, \dots, r_n\}$ show that $\{br_1, \dots, br_n\} = R\}]$
 - Every idempotent in R is central.
27. (a) If $a^m = a^{m+n}$ in a ring R , where $n \geq 1$, show that $(a^t)^2 = a^t$ for some t . [Hint: $a^{m+r} = a^{m+kn+r}$ for all $k \geq 0$.]
(b) If R is finite, show that some power of each element is an idempotent.
28. In each case find the units, the nilpotents, and the idempotents in R .
- $R = \mathbb{Z}$
 - $R = \mathbb{Z}_{24}$
 - $R = M_2(\mathbb{Z}_2)$
 - $R = \begin{bmatrix} \mathbb{R} & \mathbb{R} \\ 0 & \mathbb{R} \end{bmatrix}$
29. Let $R = \begin{bmatrix} \mathbb{Z} & X \\ 0 & \mathbb{Z} \end{bmatrix} = \left\{ \begin{bmatrix} n & x \\ 0 & m \end{bmatrix} \mid n, m \in \mathbb{Z}; x \in X \right\}$ where X is any abelian group. Show that R is a ring with the usual matrix operations and find the units, nilpotents and idempotents.
30. Show that $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is invertible in $M_2(R)$ if a and $d - ca^{-1}b$ are invertible in R . [Hint: Find p and q such that $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & 0 \\ c & 1 \end{bmatrix} \begin{bmatrix} 1 & p \\ 0 & q \end{bmatrix}$.]
31. If m is odd, show that \bar{m} is an idempotent in \mathbb{Z}_{2m} .
32. If $a^{2m} = a$ for all $a \in R$, show that $2a = 0$ for all $a \in R$.
33. A ring R is called a **boolean ring** (after George Boole) if $r^2 = r$ for all $r \in R$. Show that every boolean ring $R \neq 0$ is commutative of characteristic 2.
34. Let $R = \{X \mid X \subseteq U\}$ where U is a set. If $X \oplus Y = (X \setminus Y) \cup (Y \setminus X)$ and $XY = X \cap Y$, show that R is a boolean ring (Exercise 33).

35. Show that $\begin{bmatrix} R & R \\ 0 & R \end{bmatrix} \cong \begin{bmatrix} R & 0 \\ R & R \end{bmatrix}$ for any ring R .
36. In each case show that the given rings are not isomorphic.
 (a) \mathbb{R} and \mathbb{C} (b) \mathbb{Q} and \mathbb{R} (c) \mathbb{Z} and \mathbb{Q} (d) \mathbb{Z}_8 and $\mathbb{Z}_4 \times \mathbb{Z}_2$
37. If R and R' are rings and $\sigma : R \rightarrow R'$ is an onto mapping satisfying $\sigma(rs) = \sigma(r) \cdot \sigma(s)$ for all $r, s \in R$, show that $\sigma(1) = 1$.
38. If $R \cong S$ are rings, show that: (a) $Z(R) \cong Z(S)$; (b) $R^* \cong S^*$ (as groups).
39. Let X be an additive abelian group. A group homomorphism $\alpha : X \rightarrow X$ is called an **endomorphism** of X . Given another endomorphism $\beta : X \rightarrow X$, define the sum $\alpha + \beta : X \rightarrow X$ by $(\alpha + \beta)(x) = \alpha(x) + \beta(x)$ for all $x \in X$. Show that the set X of all endomorphisms of X is a ring using this addition, and using composition as the multiplication.
40. Write $M_2(R) = S$. Find an idempotent $e^2 = e$ in S such that $R \cong eSe$.
41. If $e^2 = e \in R$, show that $\sigma : (eRe)^* \rightarrow R^*$ is a one-to-one group homomorphism where $\sigma(a) = a + (1 - e)$ for all $a \in (eRe)^*$.
42. (a) Show that $\bar{k} \in \mathbb{Z}_n$ is nilpotent if and only if all prime divisors of n divide k .
 (b) If $n = ab$ where $\gcd(a, b) = 1$, and if $1 = xa + yb$ where $x, y \in \mathbb{Z}$, show that \overline{xa} is an idempotent in \mathbb{Z}_n .
 (c) Show that every idempotent in \mathbb{Z}_n arises as in (b). [Hint: Exercise 35 §1.2.]
43. Show that there are four nonisomorphic rings of order 4, isomorphic to one of \mathbb{Z}_4 , $\mathbb{Z}_2 \times \mathbb{Z}_2$, $L = \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in \mathbb{Z}_2 \right\}$, and a field (see Section 4.3).

3.2 INTEGRAL DOMAINS AND FIELDS

We have shown that $a \cdot 0 = 0 = 0 \cdot a$ holds for every element a in any ring. One of the most useful properties of the ring \mathbb{R} of real numbers is that the *only* way that a product can equal zero is if one of the factors is zero; that is, $ab = 0$ in \mathbb{R} implies that $a = 0$ or $b = 0$. This property is a fundamental tool for solving equations. For example, the usual method for solving the quadratic equation $x^2 - x - 12 = 0$, $x \in \mathbb{R}$, is first to factor it as $(x - 4)(x + 3) = 0$ and then conclude that $x - 4 = 0$ or $x + 3 = 0$; that is, $x = 4$ or $x = -3$. In this section we investigate rings in which $ab = 0$ implies that $a = 0$ or $b = 0$. The next theorem identifies two other equivalent conditions.

Theorem 1. *The following conditions are equivalent for a ring R .*

- (1) *If $ab = 0$ in R , then either $a = 0$ or $b = 0$.*
- (2) *If $ab = ac$ in R and $a \neq 0$, then $b = c$.*
- (3) *If $ba = ca$ in R and $a \neq 0$, then $b = c$.*

Proof. (1) \Rightarrow (2). Given (1), let $ab = ac$, where $a \neq 0$. Then $ab - ac = 0$, so $a(b - c) = 0$. As $a \neq 0$, (1) implies that $b - c = 0$; that is, $b = c$.

(2) \Rightarrow (1). Assume (2) and let $ab = 0$ in R . If $a = 0$, there is nothing to prove. If $a \neq 0$, the fact that $ab = a0 (= 0)$ gives $b = 0$ by (2). ■

The proof that (1) \Leftrightarrow (3) is analogous. ■

A ring $R \neq 0$ is called a **domain** if the conditions in Theorem 1 are satisfied. A commutative domain is called an **integral domain**.

Example 1. \mathbb{Z} is an integral domain—hence the name.

Example 2. Show that every division ring is a domain, and hence that every field is an integral domain.

Solution. Let $ab = 0$ in the division ring R ; we must show that $a = 0$ or $b = 0$. But if $a \neq 0$, then a^{-1} exists by hypothesis, so $b = 1b = a^{-1}ab = a^{-1}0 = 0$. Thus, R is a domain. \square

Example 3. Show that every subring of a division ring (a field) is a domain (integral).

Solution. Let R be a subring of a division ring D . If $ab = 0$ in R , then $ab = 0$ in D too, so $a = 0$ or $b = 0$ by Example 2. Thus, R is a domain. If D is a field, R is commutative and so is an integral domain. \square

Thus, the ring $\mathbb{Z}(i) = \{m + ni \mid m, n \in \mathbb{Z}\}$ of gaussian integers is an integral domain. In fact many interesting examples of fields and integral domains arise as subrings of \mathbb{C} . Here is an example that is actually a field.

Example 4. Write $\mathbb{Q}(\sqrt{2}) = \{r + s\sqrt{2} \mid r, s \in \mathbb{Q}\}$. Show that $\mathbb{Q}(\sqrt{2})$ is a field.

Solution. Verifying that $\mathbb{Q}(\sqrt{2})$ is a subring of \mathbb{R} is easy. To verify that it is a field, it is convenient to introduce the following notions: By analogy with \mathbb{C} , given $a = r + s\sqrt{2}$ in $\mathbb{Q}(\sqrt{2})$, define its conjugate a^* and norm $N(a)$ by

$$a^* = r - s\sqrt{2} \quad \text{and} \quad N(a) = r^2 - 2s^2.$$

Observe that $N(a) = aa^*$. Suppose now that $a \neq 0$ in $\mathbb{Q}(\sqrt{2})$. If $a = r + s\sqrt{2}$ then $s \neq 0$ because $\sqrt{2} \notin \mathbb{Q}$ (Example 3 §0.1). Hence $N(a) = r^2 - 2s^2 \neq 0$ in \mathbb{Q} . But then $\frac{1}{N(a)} \in \mathbb{Q}$, so the fact that $aa^* = N(a)$ implies that $a^{-1} = \frac{1}{N(a)}a^*$ exists in $\mathbb{Q}(\sqrt{2})$. Hence, $\mathbb{Q}(\sqrt{2})$ is a field. \square

The analogy between $\mathbb{Q}(\sqrt{2})$ and \mathbb{C} goes further: It is not difficult to verify that $(ab)^* = a^*b^*$ holds for all a and b in $\mathbb{Q}(\sqrt{2})$, and hence that $N(ab) = N(a)N(b)$. Some consequences of this are explored in Exercise 21.

The ring $\mathbb{Q}(\sqrt{2})$ in Example 4 is the result of *adjoining* an element $\sqrt{2}$ (not in \mathbb{Q}) to the field \mathbb{Q} . In this case everything is going on inside \mathbb{R} , and the resulting ring is a subring of \mathbb{R} . Similarly, the gaussian integers $\mathbb{Z}(i)$ are the result of adjoining i to \mathbb{Z} inside \mathbb{C} . This adjoining process works more generally.

For example, if R is any ring, we write $R(\omega)$ to denote all formal sums $r + sw$, where r and s are in R :

$$R(\omega) = \{r + sw \mid r, s \in R\}.$$

As in \mathbb{C} , we decree that

$$r + sw = r' + s'\omega \quad \text{if and only if} \quad r = r' \text{ and } s = s',$$

and we insist that

$$\omega^2 \in R \quad \text{and} \quad r\omega = \omega r, \quad \text{for all } r \in R.$$

Then the ring axioms determine the addition and multiplication in $R(\omega)$. Taking $\omega = i$, we obtain $\mathbb{R}(i) = \mathbb{C}$, and $\mathbb{Z}(i)$ is the ring of gaussian integers as before.

We investigate this construction, and others like it, in Section 5.2. For the present, we use it informally to construct a field of nine elements.

Example 5. Show that $\mathbb{Z}_3(\omega)$ is a field with nine elements if $\omega^2 = -1$.

Solution. Write $\mathbb{Z}_3 = \{0, 1, 2\}$. Then

$$\mathbb{Z}_3(\omega) = \{0, 1, 2, \omega, 2\omega, 1 + \omega, 1 + 2\omega, 2 + \omega, 2 + 2\omega\}.$$

For $a = r + s\omega$ in $\mathbb{Z}_3(\omega)$, write $a^* = r - s\omega$, so that $aa^* = r^2 + s^2 \in \mathbb{Z}_3$. If $a \neq 0$, then $r \neq 0$ or $s \neq 0$ holds in \mathbb{Z}_3 —by our definition of $R(\omega)$. This means that $r^2 + s^2 \neq 0$ in \mathbb{Z}_3 (in fact, $r^2 = 0$ or 1 for all r in \mathbb{Z}_3). Write $b = (r^2 + s^2)^{-1}a^*$. Then $ab = 1 = ba$ so $b = a^{-1}$. \square

We now turn to other properties of domains.

Theorem 2. The characteristic of any domain is either zero or a prime.

Proof. Let R be a domain, suppose $\text{char } R \neq 0$, say $\text{char } R = n > 0$. If n is not a prime, let $n = km$, where $1 < k < n$ and $1 < m < n$. If 1 is the unity of R , then Theorem 2 §3.1 gives $(k1)(m1) = (km)(1 \cdot 1) = n1 = 0$. Hence, $k1 = 0$ or $m1 = 0$ because R is a domain, a contradiction because $n = o(1)$. Hence, n is a prime. \blacksquare

Because $\text{char } \mathbb{Z}_n = n$ for each $n \geq 2$, Theorem 2 shows that \mathbb{Z}_n is an integral domain if and only if n is a prime, that is, if and only if \mathbb{Z}_n is a field. This also follows from Theorem 3.

Theorem 3. Every finite integral domain is a field.

Proof. Let R be a finite integral domain, say $|R| = n$, and write $R = \{r_1, r_2, \dots, r_n\}$. Given $a \neq 0$ in R , the set $aR = \{ar_1, ar_2, \dots, ar_n\}$ has distinct elements ($ar_i = ar_j$ implies $r_i = r_j$ by Theorem 1). Hence $|aR| = n$ so, since $aR \subseteq R$ and $|R| = n$, we have $aR = R$. In particular, $1 \in aR$, say $1 = ab$, $b \in R$. Because R is commutative, this shows that a is a unit. Hence, R is a field. \blacksquare

A similar argument shows that every finite domain is a division ring (Exercise 23). The reason for only considering the commutative case is a remarkable theorem first proved in 1905 by J.H.M. Wedderburn. We prove it in Section 10.4.

Wedderburn's Theorem. Every finite division ring is a field.

This theorem seems to indicate that noncommutative division rings are rare. However, an example called the quaternions has been known since 1843.

Quaternions

In the early part of the nineteenth century, the importance of the complex numbers was becoming increasingly apparent. The Irish mathematician William Rowan Hamilton gave the first modern exposition of the complex numbers in 1833. The set of complex numbers can be identified with the points in the plane, and Hamilton was looking for an analogous algebra to describe three-dimensional space. After a frustrating 10-year search, he finally realized that the algebra he sought must be four-dimensional and that the commutative law must fail. He called these new

“numbers” **quaternions** and subsequently devoted a great deal of time to them. However, their use has been limited by the great success of vector analysis.

Complex numbers have the form $a + bi$ where a and b are real and $i^2 = -1$. By analogy, the set \mathbb{H} of quaternions is defined by

$$\mathbb{H} = \{a + bi + cj + dk \mid a, b, c, d \text{ in } \mathbb{R}\}.$$

Here, as for complex numbers, we require that

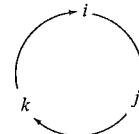
$$a + bi + cj + dk = a' + b'i + c'j + d'k \Leftrightarrow a = a', b = b', c = c', \text{ and } d = d'.^{50}$$

We also insist that each $r \in \mathbb{R}$ commutes with each of i, j , and k . With this, the multiplication in \mathbb{H} is determined by the distributive laws once the products i^2, j^2, ij, \dots are specified. These in turn follow from the equations

$$i^2 = j^2 = k^2 = ijk = -1,^{51}$$

which yield the following formulas:

$$\begin{aligned} ij &= k = -ji \\ jk &= i = -kj \\ ki &= j = -ik \end{aligned}$$



These formulas are best remembered from the diagram: The product of any two of i, j and k taken clockwise around the circle is the next one, while the product counterclockwise is the negative of the next one.⁵²

The fact that \mathbb{H} is associative can be either verified directly, or by noting that there is a concrete realization of \mathbb{H} as a subring of the ring $M_2(\mathbb{C})$ of 2×2 matrices over \mathbb{C} (Exercise 31). The ring \mathbb{C} of complex numbers is regarded as a subring of \mathbb{H} by identifying $a + bi = a + bi + 0j + 0k$.

The following example illustrates how products in \mathbb{H} are computed.

Example 6. $(3 - 4j)(2i + k) = 6i + 3k - 8ji - 4jk = 6i + 3k + 8k - 4i = 2i + 11k.$

Example 7. Show that \mathbb{R} is the center of \mathbb{H} .

Solution. If $a \in \mathbb{R}$ then $aq = qa$ for all $q \in \mathbb{H}$ because a commutes with i, j , and k . Conversely, let $q = a + bi + cj + dk$ lie in $Z(\mathbb{H})$. Then the fact that $qi = iq$ gives $-b + ai + dj - ck = -b + ai - dj + ck$. Equating coefficients gives $c = 0 = d$, so $q = a + bi$. But then $qj = jq$ implies $b = 0$, so $q = a \in \mathbb{R}$, as required. \square

If $z = a + bi$ is a complex number, we have $\bar{z} = a - bi$ and $|z|^2 = a^2 + b^2$. The analogy between \mathbb{C} and \mathbb{H} leads to a natural extension of these important notions to \mathbb{H} . Given a quaternion $q = a + bi + cj + dk$, define the **conjugate** q^* and the **norm** $N(q)$ as follows:

$$q^* = a - bi - cj - dk \quad \text{and} \quad N(q) = a^2 + b^2 + c^2 + d^2.$$

⁵⁰In other words, \mathbb{H} is a four-dimensional vector space over \mathbb{R} with basis $\{1, i, j, k\}$.

⁵¹These equations first occurred to Hamilton while he was out walking, and he was so impressed with their importance that he carved the symbols with a knife on Brougham Bridge in Dublin. The date was October 6, 1843.

⁵²In Section 2.8, the set $Q = \{\pm 1, \pm i, \pm j, \pm k\}$ was called the quaternion group.

A routine calculation establishes the following fact:

$$qq^* = N(q) = q^*q \quad \text{for every quaternion } q.$$

With this we prove

Theorem 4. *The ring \mathbb{H} is a noncommutative division ring. Moreover, if $q \neq 0$ in \mathbb{H} , then $q^{-1} = \frac{1}{N(q)}q^*$.*

Proof. \mathbb{H} is noncommutative (for example, $ij \neq ji$). If $q = a + bi + cj + dk \neq 0$ in \mathbb{H} , then one of a, b, c , or d is nonzero, so $N(q) = a^2 + b^2 + c^2 + d^2 \neq 0$ in \mathbb{R} . Since $N(q) \in \mathbb{R}$ is central in \mathbb{H} , the equations $qq^* = N(q) = q^*q$ give $q^{-1} = \frac{1}{N(q)}q^*$. ■

We mention one more fact about \mathbb{H} . It is not difficult to verify (Exercise 30) that the norm is multiplicative in the sense that

$$N(pq) = N(p)N(q), \quad \text{for all } p \text{ and } q \text{ in } \mathbb{H}.$$

This formula shows that the product $N(p) \cdot N(q)$ of two sums of four squares can itself be written as a sum of four squares. This is Lagrange's famous **four square identity**. The analogue for two squares is also true, and is a consequence of the fact that $|zw|^2 = |z|^2|w|^2$ for any complex numbers z and w .

Field of Quotients

By Example 3, every subring of a field is an integral domain. The converse also holds: Every integral domain R is isomorphic to a subring of a field F (we say R is **embedded** in F). The prototype example is \mathbb{Z} , where we regard $\mathbb{Z} \subseteq \mathbb{Q}$ by identifying the integer n with the fraction $\frac{n}{1}$. More generally, if R is any integral domain, we construct a field Q of all *fractions* or *quotients* $\frac{r}{u}$ from R and show that R can be identified with a subring of Q .

The fact that, for example, $\frac{3}{5}$ and $\frac{21}{35}$ are equal fractions must seem mysterious when it is first encountered in school, and some pupils probably are not too enlightened when the teacher points out that $\frac{21}{35} = \frac{3 \cdot 7}{5 \cdot 7}$. The reason, of course, is that a fraction such as $\frac{3}{5}$ represents a whole *class* of pairs of integers (m, n) , where $\frac{m}{n} = \frac{3}{5}$. This representation suggests that an equivalence relation is at work. Our jumping-off point is the observation that $\frac{m}{n} = \frac{m'}{n'}$ in \mathbb{Q} if and only if $mn' = m'n$.

This last equation makes sense in any integral domain R , and we use it to construct quotients $\frac{r}{u}$ from R as equivalence classes. First, we let

$$X = \{(r, u) \mid r \in R, u \in R, u \neq 0\}$$

and define a relation \equiv on X by

$$(r, u) \equiv (s, v), \quad \text{if and only if} \quad rv = su.$$

We claim that this is an equivalence on X . Clearly, $(r, u) \equiv (r, u)$ for all (r, u) in X , and $(r, u) \equiv (s, v)$ implies that $(s, v) \equiv (r, u)$. To prove transitivity, let

$$(r, u) \equiv (s, v) \quad \text{and} \quad (s, v) \equiv (t, w).$$

Then $rv = su$ and $sw = tv$, so $(rw)v = (rv)w = (su)w = u(sw) = utv$ (as R is commutative). We have $v \neq 0$ because $(s, v) \in X$, so we may cancel v in the domain R to obtain $rw = tu$; that is,

$$(r, u) \equiv (t, w).$$

Thus \equiv is an equivalence on X .

Motivated by the case $R = \mathbb{Z}$, we define the quotient $\frac{r}{u}$ to be the equivalence class $[(r, u)]$ of the pair (r, u) in X . More precisely, we write

$$\frac{r}{u} = [(r, u)].$$

Now we invoke Theorem 1 §0.4 that $[(r, u)] = [(s, v)]$ if and only if $(r, u) \equiv (s, v)$. In our quotient notation, this extends the familiar fact about rational fractions:

$$\frac{r}{u} = \frac{s}{v} \quad \text{if and only if} \quad rv = su. \quad (*)$$

Moreover, this condition implies another useful property of rational fractions:

$$\frac{r}{u} = \frac{vr}{vu}, \quad \text{for all } v \neq 0 \text{ in } R. \quad (**)$$

So we have created the quotients we wanted.

Now let Q denote the set of all these quotients; that is,

$$Q = \left\{ \frac{r}{u} \mid r, u \text{ in } R \text{ and } u \neq 0 \right\}.$$

Our objective is to make Q into a field. Once again motivated by \mathbb{Q} , we define addition and multiplication in Q by

$$\frac{r}{u} + \frac{s}{v} = \frac{rv + su}{uv} \quad \text{and} \quad \frac{r}{u} \cdot \frac{s}{v} = \frac{rs}{uv}$$

where $rv + su$, rs and uv on the right-hand side of each equation are computed in R . Note that $uv \neq 0$ — R is a domain, so these are legitimate quotients in Q .

Because these quotients are equivalence classes, we must show that addition and multiplication are well defined by these formulas. We do it for addition: If $\frac{r}{u} = \frac{r'}{u'}$ and $\frac{s}{v} = \frac{s'}{v'}$, we must show that $\frac{rv + su}{uv} = \frac{r'v' + s'u'}{u'v'}$. We have $ru' = r'u$ and $sv' = s'v$ by $(*)$, and we must show that $uv(r'v' + s'u') = u'v'(rv + su)$, again by $(*)$. Compute

$$uv(r'v' + s'u') = (r'u)vv' + (s'v)uu' = (ru')vv' + (sv')uu' = u'v'(rv + su)$$

as required. The verification that multiplication is well defined is left to the reader.

With this, we can show that Q really is a field. Most of this will also be left to the reader; we verify the associative law of addition:

$$\begin{aligned} \frac{r}{u} + \left(\frac{s}{v} + \frac{t}{w} \right) &= \frac{r}{u} + \left(\frac{sw + tv}{vw} \right) = \frac{r(vw) + (sw + tv)u}{u(vw)} \\ &= \frac{(rv + su)w + t(uv)}{(uv)w} = \left(\frac{rv + su}{uv} \right) + \frac{t}{w} \\ &= \left(\frac{r}{u} + \frac{s}{v} \right) + \frac{t}{w}. \end{aligned}$$

Similar calculations show that Q is a commutative ring where, if $u \neq 0$ in R , the zero is $\frac{0}{1} = \frac{0}{u}$, the unity is $\frac{1}{1} = \frac{u}{u}$, and the negative of $\frac{r}{u}$ is $\frac{-r}{u}$. Moreover, if $\frac{r}{u}$ is nonzero in Q , then $r \neq 0$, so $\frac{u}{r} \in Q$. Then $(**)$ gives

$$\frac{r}{u} \cdot \frac{u}{r} = \frac{ru}{ru} = \frac{1}{1} \quad \text{is the unity of } Q.$$

Hence $(\frac{r}{u})^{-1} = \frac{u}{r}$, and we have proved that Q is a field.

Finally, we can easily verify that $R' = \{\frac{r}{1} \mid r \in R\}$ is a subring of Q . Let $\sigma : R \rightarrow R'$ be defined by $\sigma(r) = \frac{r}{1}$ for all $r \in R$. Then σ is clearly onto, and it is one-to-one because $\frac{r}{1} = \frac{s}{1}$ implies that $r = s$ by (*). Moreover, σ is a ring isomorphism because

$$\frac{r}{1} + \frac{s}{1} = \frac{r+s}{1} \quad \text{and} \quad \frac{r}{1} \cdot \frac{s}{1} = \frac{rs}{1}.$$

Hence $R \cong R'$. Customary practice is to identify $R = R'$ by taking $r = \frac{r}{1}$ for all $r \in R$ (as in $\mathbb{Z} \subseteq \mathbb{Q}$), and so to regard R as an actual subring of Q .

Theorem 5. Embedding Theorem. If R is an integral domain, there is a field Q consisting of quotients $\frac{r}{u}$, where r and $u \neq 0$ are elements of R . By identifying $r = \frac{r}{1}$ for all $r \in R$ we may (and do) regard R as a subring of Q . In that case, every $u \neq 0$ in R has an inverse in Q , and each quotient in Q has the form $\frac{r}{u} = ru^{-1}$, where r and $u \neq 0$ are in R .

Proof. Only the last sentence remains to be proved. We have

$$\frac{r}{u} = \frac{r}{1} \cdot \frac{1}{u} = \frac{r}{1} \cdot \left(\frac{u}{1}\right)^{-1}$$

which becomes ru^{-1} if we identify $r = \frac{r}{1}$ for all $r \in R$. ■

The field Q constructed in Theorem 5 is called the **field of quotients** of the integral domain R .

The construction of the field Q of quotients of an integral domain R depends heavily on the fact that R is commutative. This dependence is in fact essential, because there exist noncommutative domains that cannot be embedded in a division ring. The first such example was discovered in 1937 by the Russian mathematician Anatoly Ivanovich Mal'cev.⁵³ On the other hand, a wide class of noncommutative domains can be embedded in a division ring of *right* quotients. These are called **right Ore-domains**, after Oystein Ore who first discussed them in 1931.

Exercises 3.2

Throughout these exercises R denotes a ring unless otherwise specified.

1. Find all the roots of $x^2 + 3x - 4$ in
 - (a) \mathbb{Z}
 - (b) \mathbb{Z}_6
 - (c) \mathbb{Z}_4
2. If p is a prime, let $\mathbb{Z}_{(p)} = \{\frac{n}{m} \in \mathbb{Q} \mid p \text{ does not divide } m\}$. Show that this is an integral domain and find all the units.
3. Determine all idempotents and nilpotents in a domain.
4. Is $R \times S$ ever a domain? Support your answer.
5. Show that $M_n(R)$ is never a domain if $n \geq 2$.
6. If $a^2 = b^2$ and $a^3 = b^3$ in a domain, show that $a = b$. Now do it if $a^m = b^m$ and $a^n = b^n$ where $\gcd(m, n) = 1$. [Hint: $1 = xm + yn$, where $x, y \in \mathbb{Z}$.]

⁵³Mal'cev, A.I., Groups and other algebraic systems, in *Mathematics: Its Contents and Meaning*, Vol. 3, Cambridge MA: MIT Press, 1963.

7. Suppose that R has no nonzero nilpotent elements (for example, a domain). If $ab = 0$ in R , show that $ba = 0$.
8. Show that a ring R is a division ring if and only if, for each nonzero $a \in R$, there is a unique element $b \in R$ such that $aba = a$.
9. Find a finite field in which $a^2 + b^2 = 0$ implies that $a = b = 0$, and find another in which this is not true.
10. If $F = \{0, 1, a, b\}$ is a field, fill in the addition and multiplication tables for F .
11. If F is a field and $|F| = q$, show that $a^q = a$ for all $a \in F$. [Hint: Lagrange.]
12. Show that the characteristic of a finite field must be a prime.
13. If F is a field and $|F| = p$, where p is a prime, show that $F \cong \mathbb{Z}_p$.
14. Show that there is no field of order 6. [Hint: Lagrange's theorem.]
15. Show that the center of a division ring is a field.
16. Let K be a subring of a field F . Call K a **subfield** of F if it is a field using the operations of F .
 - (a) Show that K is a subfield of F if and only if $0 \neq a \in K$ implies that $a^{-1} \in K$.
 - (b) If $|F| = 8$ and K is a subfield, show that $K = F$ or $K = \{0, 1\}$. [Hint: Lagrange.]
 - (c) What happens if $|F| = 16$ in (b)?
17. Show that $\mathbb{Q}(i) = \{r + si \mid r, s \in \mathbb{Q}\}$ is a subfield of \mathbb{C} .
18. (a) Show that $\mathbb{Q}(\sqrt{5}i) = \{r + s\sqrt{5}i \mid r, s \in \mathbb{Q}\}$ is a subfield of \mathbb{C} . [Hint: Example 4.]
 (b) Show that $\mathbb{Z}(\sqrt{5}i) = \{n + m\sqrt{5}i \mid n, m \in \mathbb{Z}\}$ is a subring of \mathbb{C} and find the units.
 [Hint: Example 14 §3.1.]
19. Show that $\mathbb{Q}(\sqrt{2})$ is the smallest subfield of \mathbb{R} that contains $\sqrt{2}$.
20. Show that $\mathbb{Z}(\sqrt{2}) = \{n + m\sqrt{2} \mid n, m \in \mathbb{Z}\}$ is a subring of \mathbb{C} and find 10 units (in fact there are infinitely many). [Hint: Example 4.]
21. Let $w \in \mathbb{C}$ satisfy $w^2 \in \mathbb{Z}$, but $w \notin \mathbb{Q}$, and define $\mathbb{Z}(w) = \{n + mw \mid n, m \in \mathbb{Z}\}$. If $r = n + mw$ is in $\mathbb{Z}(w)$ write $r^* = n - mw$ and $N(r) = n^2 - w^2m^2$. Show that:
 - (a) $\mathbb{Z}(w)$ is an integral domain.
 - (b) $n + mw = n' + m'w$ in $\mathbb{Z}(w)$ if and only if $n = n'$ and $m = m'$.
 - (c) $r^{**} = r$, $(rs)^* = r^*s^*$ and $(pr + qs)^* = pr^* + qs^*$ for all $p, q \in \mathbb{Z}$ and $r, s \in \mathbb{Z}(w)$.
 - (d) $N(r) = rr^*$ and $N(rs) = N(r)N(s)$ for all $r, s \in \mathbb{Z}(w)$.
 - (e) $r \in \mathbb{Z}(w)$ is a unit if and only if $N(r) = \pm 1$.
22. If R is a ring, show that R is an integral domain if and only if it satisfies the condition: $ab = ca$, $a \neq 0$, implies that $b = c$.
23. Show that a finite domain is a division ring (a field by Wedderburn's theorem).
24. Recall that the binomial coefficient is defined by $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ for $0 \leq r \leq n$.
 - (a) If p is a prime, show that $p \mid \binom{p}{r}$ for $1 \leq r \leq p-1$. [Hint: For $1 \leq r \leq n-1$, show that $\binom{n}{r} = \frac{n}{r} \binom{n-1}{r-1}$.]
 - (b) If $ab = ba$ in a ring of characteristic p , show that $(a+b)^p = a^p + b^p$.
 - (c) Let F be a finite field of characteristic p (p a prime). If $\sigma : F \rightarrow F$ is defined by $\sigma(a) = a^p$, show that σ is an automorphism of F (the **Frobenius automorphism**).
25. Let R be an integral domain and let $Q \supseteq R$ be the field of quotients. If $\sigma : R \rightarrow R$ is an automorphism, show that there is a unique automorphism $\bar{\sigma} : Q \rightarrow Q$ that satisfies $\bar{\sigma}(r) = \sigma(r)$ for all $r \in R$.
26. Show that the multiplication in (the construction of) the field of quotients of an integral domain:
 - (a) is well defined;
 - (b) is associative;
 - (c) satisfies the distributive laws.

27. If R is an integral domain, show that the field of quotients Q in Theorem 5 is the smallest field containing R in the following sense: If $R \subseteq F$, where F is a field, show that F has a subfield K such that $R \subseteq K$ and $K \cong Q$.
28. Let R be a commutative ring and call $u \in R$ a nonzero-divisor if $ur = 0$, $r \in R$ implies $r = 0$. Let $U \subseteq R$ be a set of nonzero-divisors in R such that $1 \in U$, and $ab \in U$ whenever $a, b \in U$. Generalize Theorem 5 by showing that a ring of quotients $Q = \left\{ \frac{r}{u} \mid r \in R, u \in U \right\}$ exists. Show further that R can be regarded as a subring of Q and, in this case, that each element of U is a unit in Q and $Q = \{ru^{-1} \mid r \in R, u \in U\}$.
29. If R is a ring, recall the definition of the ring $R(\omega)$ preceding Example 5 where $\omega^2 = -1$ and $r\omega = \omega r$ for all $r \in R$.
- Is $C(\omega)$ a field? What about $Z_5(\omega)$? $Z_7(\omega)$?
 - If R is commutative, show that $R(\omega)^* = \{r + sw \mid r^2 + s^2 \in R^*\}$.
 - If p is a prime and $p \equiv 3 \pmod{4}$, show that $Z_p(\omega)$ is a field of order p^2 . [Hint: Corollary to Theorem 8 §1.3.]
 - If R is an integral domain in which $2 \neq 0$, show that $R(\omega)$ has no nonzero nilpotents.
 - If R is an integral domain in which $2 \in R^*$, show that the idempotents in $R(\omega)$ are $0, 1$, and $\frac{1}{2} + sw$, where $(2s)^2 = -1$.
 - Show that $R(\omega) \cong \left\{ \begin{bmatrix} r & -s \\ s & r \end{bmatrix} \mid r, s \in R \right\}$, a subring of $M_2(R)$.
30. Let p and q denote quaternions and let $a, b \in \mathbb{R}$. Show that
- $(q^*)^* = q$
 - $(ap + bq)^* = ap^* + bq^*$
 - $N(q) = qq^* = q^*q$
 - $(pq)^* = q^*p^*$ [Hint: First show that $(iq)^* = -q^*i$, $(jq)^* = -q^*j$, and $(kq)^* = -q^*k$, and then use (b).]
 - $N(pq) = N(p)N(q)$ [Hint: (c) and (d).]
31. Write $1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $i = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$, $j = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, and $k = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$ in $M_2(\mathbb{C})$. Show that
- $i^2 = j^2 = k^2 = ijk = -1$.
 - $a + bi + cj + dk = \begin{bmatrix} a + bi & c + di \\ -c + di & a - bi \end{bmatrix}$ for all a, b, c , and d in \mathbb{R} .
 - $a + bi + cj + dk = a' + b'i + c'j + d'k$ if and only if $a = a', b = b', c = c'$, and $d = d'$.
 - $a\hat{i} = \hat{i}a$, $a\hat{j} = \hat{j}a$, and $a\hat{k} = \hat{k}a$ for all $a \in \mathbb{R}$.
 - \mathbb{H} is isomorphic to $\{a + bi + cj + dk \mid a, b, c, d \in \mathbb{R}\}$.
32. If R is commutative and $\mathbb{H}(R) = \{a + bi + cj + dk \mid a, b, c, d \in R\}$, we declare that
- $$a + bi + cj + dk = a' + b'i + c'j + d'k$$
- if and only if $a = a', b = b', c = c'$, and $d = d'$. As for the quaternions, the addition and multiplication in $\mathbb{H}(R)$ are determined by the ring axioms, the conditions that $i^2 = j^2 = k^2 = ijk = -1$, and the conditions $ai = ia$, $aj = ja$, and $ak = ka$ for all $a \in R$. If $q = a + bi + cj + dk$ in $\mathbb{H}(R)$, define $q^* = a - bi - cj - dk$, and $N(q) = a^2 + b^2 + c^2 + d^2$.
- Show that q is a unit in $\mathbb{H}(R)$ if and only if $N(q)$ is a unit in R .
 - Show that $\mathbb{H}(R)$ is a division ring if and only if R is a field and $a^2 + b^2 + c^2 + d^2 = 0$ in R implies that $a = b = c = d = 0$. Is $\mathbb{H}(R)$ a division ring if $R = \mathbb{C}$, \mathbb{Z}_2 , \mathbb{Z}_3 , \mathbb{Z}_5 , \mathbb{Z}_7 , or \mathbb{Z}_{11} ?

- (c) Let $A_2(R) = \{r \in R \mid 2r = 0\}$. Show that $Z[\mathbb{H}(R)] = \{a + si + tj + uk \mid a \in R, s, t, u \in A_2(R)\}$. Describe $Z[\mathbb{H}(\mathbb{Z}_6)]$. Show that $\mathbb{H}(R)$ is commutative if and only if R has characteristic 2.
- (d) Show that $q^2 - 2aq + N(q) = 0$ for all $q = a + bi + cj + dk$ in $\mathbb{H}(R)$.

3.3 IDEALS AND FACTOR RINGS

Let R be a ring and let A be an additive subgroup of R . Then A is normal in the (abelian) additive group $(R, +)$, so we obtain the additive factor group

$$R/A = \{r + A \mid r \in R\}$$

where the (additive) cosets are defined by $r + A = \{r + a \mid a \in A\}$. The essential features of the arithmetic in R/A are collected in Lemma 1 for reference; of course, they are just translations of the same properties for multiplicative groups.

Lemma 1. *Let A be an additive subgroup of a ring R and let $r, s \in R$. The following assertions are valid in the factor group R/A .*

- (1) $r + A = s + A$ if and only if $r - s \in A$.
- (2) $(r + A) + (s + A) = (r + s) + A$.
- (3) $0 + A = A$ is the (additive) unity of R/A .
- (4) $-(r + A) = -r + A$ is the (additive) inverse of $r + A$.
- (5) $k(r + A) = kr + A$ for all $k \in \mathbb{Z}$.

In our construction of $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ in Section 1.3, the cosets were written as $\bar{k} = k + n\mathbb{Z}$, $k \in \mathbb{Z}$, and we turned \mathbb{Z}_n into a ring via the multiplication $\bar{k}\bar{m} = \bar{km}$. If A is any additive subgroup of a ring R , this suggests defining multiplication in R/A by $(r + A)(s + A) = rs + A$. However, this multiplication is well defined only for rather special subgroups A . To describe them we adopt the following notation: For any element a in R , write

$$Ra = \{ra \mid r \in R\} \quad \text{and} \quad aR = \{ar \mid r \in R\}.$$

Lemma 2. *Let A be an additive subgroup of a ring R . The following are equivalent:*

- (1) *The multiplication $(r + A)(s + A) = rs + A$ is well defined on R/A .*
- (2) *$Ra \subseteq A$ and $aR \subseteq A$ for every a in A .*

Proof. (1) \Rightarrow (2). Note first that (1) turns R/A into a ring. Hence if $r \in R$ and $a \in A$ then, using (1) and Theorem 1 §3.1, we obtain

$$ra + A = (r + A)(a + A) = (r + A)(0 + A) = r0 + A = 0 + A = A.$$

This implies that $ra \in A$, so $Ra \subseteq A$. Similarly, $aR \subseteq A$.

(2) \Rightarrow (1). If $r + A = r' + A$ and $s + A = s' + A$, we must show that $rs + A = r's' + A$. We have $r - r' \in A$ and $s - s' \in A$, so

$$rs - r's' = r(s - s') + (r - r')s' \in R(s - s') + (r - r')R \subseteq A$$

by (2). Hence $rs + A = r's' + A$, as required. ■

An additive subgroup A of a ring R is called an **ideal**⁵⁴ of R if

$$Ra \subseteq A \text{ and } aR \subseteq A \text{ for every } a \in A;$$

that is, if every multiple of an element of A is again in A .

Theorem 1. *Let A be an ideal of the ring R . Then the additive factor group R/A becomes a ring with the multiplication $(r+A)(s+A) = rs+A$. The unity of R/A is $1+A$, and R/A is commutative if R is commutative.*

Proof. Because A is an additive subgroup, R/A is an additive abelian group. The multiplication is well defined by Lemma 2. Verification that it is associative, that $1+A$ is the unity, and that the distributive laws hold is left to the reader, along with the proof that R/A is commutative if R is (Exercise 3). ■

If A is an ideal of a ring R , the ring R/A in Theorem 1 is called the **factor ring** of R by A . This definition should be compared to the definition of factor groups in Section 2.9. Clearly, ideals play a role in ring theory analogous to normal subgroups in group theory, each yielding the construction of a factor structure using cosets. Note, however, that although normal subgroups of a group are certainly subgroups, most ideals are not subrings. It is true that ideals of R are closed under multiplication (and so are general subrings) but, as the next theorem shows, the only ideal that contains the unity of R is R itself.

Theorem 2. *The following are equivalent for an ideal A of a ring R .*

- (1) $1 \in A$.
- (2) A contains a unit.
- (3) $A = R$.

Proof. (1) \Rightarrow (2) and (3) \Rightarrow (1) are obvious. If $u \in A$ is a unit then $1 = u^{-1}u \in A$ because A is an ideal. Hence, $r = r \cdot 1 \in A$ for all $r \in R$, proving (2) \Rightarrow (3). ■

Example 1. If R is any ring, $\{0\}$ and R are ideals of R , and the factor rings are $R/\{0\} \cong R$ and $R/R \cong \{R\}$ —the zero ring with one element. The ideal $0 = \{0\}$ is called the **zero ideal** of R , and any ideal $A \neq R$ is called a **proper ideal** of R .

Example 2. If $n \geq 0$, then $n\mathbb{Z}$ is an ideal of \mathbb{Z} and $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$ if $n \geq 2$.

Note that every additive subgroup of \mathbb{Z} has the form $n\mathbb{Z}$ for some $n \geq 0$, so every additive subgroup of \mathbb{Z} is an ideal. In fact, \mathbb{Z} and \mathbb{Z}_n , $n \geq 2$, are the only nonzero rings having this property.

Example 3. If $a \in Z(R)$, show that $Ra = aR$, and that this is an ideal of R called the **principal ideal generated by a** .⁵⁵

Solution. First, Ra is an additive subgroup of R because $ra + sa = (r+s)a$, $0 = 0a$, and $-(ra) = (-r)a$. If $x = sa \in Ra$ and $r \in R$, then $rx = r(sa) = (rs)a \in Ra$. Note

⁵⁴In the nineteenth century it was observed that the prime factorization theorem for the ring \mathbb{Z} of rational integers did not extend to certain subrings of \mathbb{C} . Ernst Eduard Kummer showed that unique factorization was achieved for what he called *ideal numbers*. The term *ideal* was first used by Richard Dedekind who realized that the *ideal numbers* could best be described as ideals in the modern sense.

⁵⁵In a commutative ring R , the ideal Ra of R is often denoted $\langle a \rangle$.

that we have not yet used the fact that $a \in Z(R)$. But this is needed to show that $xr = (sa)r = sra \in Ra$ for all $r \in R$. Hence, Ra is an ideal. Clearly $Ra = aR$. \square

Note that, if $a \in Z(R)$, the ideal $\langle a \rangle = Ra = aR$ in Example 3 contains a and is contained in every ideal of R that contains A . Hence, it is the *smallest* ideal of R containing a . If $a \notin Z(R)$, the description of this smallest ideal containing a is more complex (see Exercise 27).

Example 4. If $a \in Z(R)$, show that $\text{ann}(a) = \{r \in R | ra = 0\}$ is an ideal of R , called the **annihilator** of a .

Solution. The set $\text{ann}(a)$ is an additive subgroup because $0a = 0$, and $ra = sa = 0$ implies that $(r+s)a = 0 = (-r)a$. If $ra = 0$ and $t \in R$, then $(tr)a = t(ra) = 0$, and $(rt)a = rat = 0$ because $a \in Z(R)$. Hence, $tr \in \text{ann}(a)$ and $rt \in \text{ann}(a)$. \square

An ideal A of a ring R is a general ring but it contains the unity of R only if $A = R$ by Theorem 2. However, if $e^2 = e$ is a central idempotent in R then $A = Re$ is an ideal by Example 3, and e is the unity of A by Theorem 6 §3.1 (in fact $Re = eRe$). This observation has a converse that we need later.

Example 5. Let A be an ideal of a ring R , and assume that A is a ring with unity e . Show that e is a central idempotent of R , and that $A = eRe$.

Solution. Clearly $e^2 = e$ because e is the unity of A . To show that e is central, let $r \in R$ and write $a = er - ere$. Then $a \in A$ so, since e is the unity of A , $a = ae = ere - ere^2 = 0$. Hence $er = ere$, and a similar argument shows that $re = ere$. Thus $er = ere = re$ for all $r \in R$; that is e is central. Finally, if $a \in A$, then $a = ae \in Re$, so $A \subseteq Re$. Since $Re \subseteq A$ because $e \in A$, we have $A = Re$. Finally, $Re = eRe$ because e is a central idempotent. \square

Example 6 illustrates how to carry out computations in a factor ring.

Example 6. Let $R = \mathbb{Z}(i)$ be the ring of gaussian integers and let $A = (2+i)\mathbb{Z}$ denote the ideal of all multiples of $2+i$. Describe the cosets in R/A .

Solution. A typical coset x in R/A has the form $x = (m+ni) + A$, where $m, n \in \mathbb{Z}$. Since $2+i \in A$, we have $i+A = -2+A$. Hence, $x = (m-2n)+A$ in R/A ; that is,

$$x = k+A, \quad \text{for some } k \in \mathbb{Z}.$$

This simplifies even further: Note that $5 = (2+i)(2-i) \in A$, so $5+A = 0+A$. Thus, if $k = 5q+r$, $0 \leq r \leq 4$, we get $x = k+A = (5+A)(q+A) + (r+A) = r+A$ in the ring R/A . Hence,

$$R/A = \{0+A, 1+A, 2+A, 3+A, 4+A\}.$$

We claim that these five cosets are distinct. Suppose that $r+A = s+A$, where $0 \leq s \leq r \leq 4$. Then $r-s \in A$, say $r-s = (2+i)(a+bi)$ for some $a, b \in \mathbb{Z}$. Taking absolute values gives $(r-s)^2 = 5(a^2+b^2)$. As $(r-s)^2$ is 0, 1, 4, 9, or 16, the only possibility is $r=s$. Thus $|R/A| = 5$. Note, finally, that $R/A \cong \mathbb{Z}_5$ is a field by Theorem 7 §3.1. \square

An ideal P of a commutative ring R is called a **prime ideal** if $P \neq R$ and P has the following property:

$$\text{If } rs \in P, \text{ then } r \in P \text{ or } s \in P.$$

Recall that a commutative ring R is an integral domain if and only if $rs = 0$ implies $r = 0$ or $s = 0$, that is if and only if 0 is a prime ideal in R . The following characterization of prime ideals is a basic fact in the theory of commutative rings.

Theorem 3. If R is a commutative ring, an ideal $P \neq R$ of R is a prime ideal if and only if R/P is an integral domain.

Proof. If R/P is an integral domain and $rs \in P$, then $(r + P)(s + P) = rs + P = P$ is the zero of R/P , so either $r + P = P$ or $s + P = P$. Hence $r \in P$ or $s \in P$, so P is a prime ideal. Conversely, if P is a prime ideal, let $(r + P)(s + P) = P$, the zero of R/P ; that is $rs + P = P$. Hence, $rs \in P$, so $r \in P$ or $s \in P$ because P is a prime ideal. Thus $r + P = P$ or $s + P = P$, proving that R/P is a domain. It is commutative because R is commutative. ■

Example 7. If $n \geq 2$ in \mathbb{Z} , show that $n\mathbb{Z}$ is a prime ideal if and only if n is a prime.

Solution. Here, $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$, which is an integral domain if and only if n is a prime (Theorem 7 §1.3). Hence, Theorem 3 applies. □

We now describe all the ideals of a factor ring R/A in terms of the ideals of R which contain A .

Theorem 4. Let A be an ideal of a ring R .

- (1) If B is an ideal of R with $A \subseteq B$ then $B/A = \{b + A \mid b \in B\}$ is an ideal of R/A .
- (2) If \mathcal{B} is any ideal of R/A then $\mathcal{B} = B/A$ for some (unique) ideal B of R with $A \subseteq B$. In fact, $B = \{b \in R \mid b + A \in \mathcal{B}\}$.
- (3) If B and B_1 are ideals of R that contain A , then

$$B \subseteq B_1 \quad \text{if and only if} \quad B/A \subseteq B_1/A.$$

Proof. (1) This is a routine verification that we leave to the reader.

(2) Given an ideal $\mathcal{B} \subseteq R/A$, let $B = \{b \in R \mid b + A \in \mathcal{B}\}$. Then B is an ideal of R (verify), and we have $A \subseteq B$ because $a + A = 0 + A \in \mathcal{B}$ for all $a \in A$. Hence it remains to show that $\mathcal{B} = B/A$. We have $\mathcal{B} \subseteq B/A$ because $r + A \in \mathcal{B}$ implies that $r \in B$, hence $r + A \in B/A$. Conversely, if $r + A \in B/A$ then $r + A = b + A$ for some $b \in B$. But $b + A \in \mathcal{B}$ because $b \in B$, that is $r + A \in \mathcal{B}$, and we have shown that $B/A \subseteq \mathcal{B}$. Hence $B/A = \mathcal{B}$, as required.

(3) If $B \subseteq B_1$, it is clear that $B/A \subseteq B_1/A$. For the converse, assume that $B/A \subseteq B_1/A$, and let $b \in B$. Then $b + A \in B_1/A$, say $b + A = b_1 + A$ for some $b_1 \in B_1$. Hence, $b \in b_1 + A \subseteq B_1$ because $A \subseteq B_1$, so $B \subseteq B_1$ as required. ■

Simple Rings

By analogy with groups, a ring R is called a **simple ring** if $R \neq 0$ and the only ideals of R are 0 and R .

Example 8. Show that every division ring is simple.

Solution. Let $A \neq 0$ be an ideal in a division ring R . If $0 \neq r \in A$, then r is a unit (because R is a division ring), so $A = R$ by Theorem 2. □

There are simple rings that are not division rings (Theorem 7 below), but such rings must be noncommutative by the next result.

Theorem 5. *If R is commutative, then R is simple if and only if it is a field.*

Proof. Every field is simple by Example 8. Conversely, if R is simple and commutative, let $0 \neq a \in R$. Then $Ra = \{ra \mid r \in R\}$ is an ideal of R (by Example 3). Because $Ra \neq 0$ (as $a \in Ra$), the simplicity of R shows that $Ra = R$. Thus $1 \in Ra$, so $1 = ba$ for some $b \in R$. Hence, a is a unit in R , so R is a field, as required. ■

The simple rings are closely related to the following class of ideals. An ideal M in a ring R is called a **maximal ideal** of R if $M \neq R$ and the only ideals A of R such that $M \subseteq A \subseteq R$ are $A = M$ and $A = R$.

Theorem 6. *Let A be an ideal of a ring R . Then A is maximal in R if and only if R/A is a simple ring.*

Proof. Assume that A is maximal and (using Theorem 4) let B/A be a nonzero ideal of R/A , where B is an ideal of R with $A \subseteq B$. Since $B/A \neq 0$, let $0 \neq b + A \in B/A$ where $b \in B$. Then $b \in B$ but $b \notin A$, so $A \neq B$. Thus, $B = R$ by the maximality of A , hence $B/A = R/A$. This shows that R/A is simple.

Conversely, if R/A is simple, let $A \subseteq B \subseteq R$ where $B \neq A$ is an ideal of R . Then B/A is an ideal of R/A by Theorem 4, and $B/A \neq 0$ because $B \neq A$. Hence, $B/A = R/A$ by the simplicity of R/A , and so $B = R$ by (3) of Theorem 4. This shows that A is a maximal ideal of R . ■

Combining Theorems 5 and 6 gives

Corollary 1. *If R is a commutative ring, an ideal A of R is maximal if and only if R/A is a field.*

The fact that every field is an integral domain, together with Theorem 3, gives

Corollary 2. *Every maximal ideal of a commutative ring is a prime ideal.*

Note that the converse of Corollary 2 is false: In the ring \mathbb{Z} of integers, the zero ideal is prime by Theorem 3, but it is not maximal by Corollary 1 because \mathbb{Z} is an integral domain that is not a field.

We conclude this section by constructing some simple rings other than division rings. In fact, we verify that $M_n(R)$ is a simple ring if R is a division ring. The proof requires some preliminary remarks about certain special matrices.

Let R be a ring and let $n \geq 1$ be a fixed integer. If $1 \leq i, j \leq n$, let E_{ij} denote the $n \times n$ matrix with (i, j) -entry 1 and all other entries 0. The matrices E_{ij} where $1 \leq i, j \leq n$, are called **matrix units** in $M_n(R)$. Thus, the matrix units in $M_2(R)$ are

$$E_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad E_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad E_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \text{and} \quad E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

In general there are n^2 matrix units in $M_n(R)$.

If A is an ideal of a ring R , it is easy to show that $M_n(A)$ is an ideal of $M_n(R)$. The converse holds too. To prove it, define the **scalar product** rA , $r \in R$, by:

If $A = [a_{ij}]$ is any matrix and $r \in R$, define $rA = [ra_{ij}]$.

Thus, rE_{ij} is the matrix with (i, j) -entry r and all other entries 0. Hence, every matrix is a “linear combination” of the matrix units E_{ij} , where $1 \leq i, j \leq n$:

$$[a_{ij}] = \sum_{i,j} a_{ij} E_{ij}. \quad (*)$$

Moreover, matrix multiplication gives the following useful formula for $r, s \in R$:

$$(rE_{ij})(sE_{kl}) = \begin{cases} rsE_{il}, & \text{if } j = k, \\ 0, & \text{if } j \neq k. \end{cases} \quad (**)$$

Lemma 3. Every ideal of $M_n(R)$ has the form $M_n(A)$ for some ideal A of R .

Proof. If \mathcal{A} is an ideal of $M_n(R)$, let $A = \{a \in R \mid aE_{11} \in \mathcal{A}\}$. Then A is an ideal of R (verify), and we show that $\mathcal{A} = M_n(A)$. To see that $M_n(A) \subseteq \mathcal{A}$, let $X \in M_n(A)$, say $X = [a_{pq}]$ where $a_{pq} \in A$ for $1 \leq p, q \leq n$. Then $(**)$ gives $a_{pq}E_{pq} = E_{p1}(a_{pq}E_{11})E_{1q} \in \mathcal{A}$ because $a_{pq}E_{11} \in \mathcal{A}$. Then, using condition $(*)$, we have $X = [a_{pq}] = \sum_{p,q} a_{pq}E_{pq} \in \mathcal{A}$, proving that $M_n(A) \subseteq \mathcal{A}$.

Conversely, let $B = [b_{ij}] \in \mathcal{A}$. We must show that $b_{pq} \in A$ for all p and q , that is $b_{pq}E_{11} \in \mathcal{A}$. Since $B = \sum_{i,j} b_{ij}E_{ij}$ by $(*)$, we compute

$$E_{1p}BE_{q1} = E_{1p}(\sum_{i,j} b_{ij}E_{ij})E_{q1} = \sum_{i,j} E_{1p}(b_{ij}E_{ij})E_{q1} = b_{pq}E_{11}$$

by $(**)$. Hence, $b_{pq}E_{11} = E_{1p}BE_{q1} \in \mathcal{A}$ because $B \in \mathcal{A}$ and \mathcal{A} is an ideal. ■

Theorem 7. If R is a ring then $M_n(R)$ is simple if and only if R is simple.

Proof. Assume that R is simple. If \mathcal{A} is an ideal of $M_n(R)$ then Lemma 3 shows that $\mathcal{A} = M_n(A)$ for some ideal A of R . Hence, A is 0 or R because R is simple, whence $\mathcal{A} = M_n(0) = 0$ or $\mathcal{A} = M_n(R)$. This shows that $M_n(R)$ is simple. The converse is proved similarly, and we leave it to the reader. ■

Corollary. If R is a division ring then $M_n(R)$ is simple.

Thus, for example, $M_2(\mathbb{Z}_2)$ is a simple noncommutative ring that is not a division ring. In fact it can be shown to be the smallest such ring.

Part of the importance of Theorem 7 is that it gives half of another theorem of Wedderburn: a “sufficiently small” ring S is simple if and only if $S \cong M_n(D)$ for some division ring D and some $n \geq 1$. We explain what “sufficiently small” means in Section 11.2 (being finite is enough), and give a proof of a more general version of the theorem.

Richard Dedekind (1831–1916). Richard Dedekind, the son of a law professor, was born in Brunswick, Germany, the birthplace of Gauss. He obtained his Ph.D. at Göttingen at the age of 21 and was Gauss' last student. After a stay in Zurich, he returned to the technical high school at Brunswick, where he remained for 50 years. He never married and lived with his sister until his death.

Dedekind had wide mathematical interests. He became disturbed by the lack of a precise foundation for the set \mathbb{R} of real numbers, and he filled this gap with his now-famous *Dedekind cuts* in a paper in 1872. His work in algebra also was of first importance. He lectured on group theory before Jordan, and stated the Peano axioms before Peano. Dedekind was one of the first to understand Galois theory and made fundamental contributions to the theory of group characters. He also extended earlier work of Kummer. The unique factorization of integers into primes is not true of elements in other integral

domains, and Kummer had shown that the uniqueness could be retrieved if certain *ideal numbers* were used. Dedekind coined the term *ideal* and studied integral domains (now called *Dedekind domains*) where all ideals factor uniquely as a product of prime ideals. This work influenced Emmy Noether, thereby changing the course of modern algebra. Dedekind also did pioneering work in the theory of rings, groups, and fields and has been called (by Morris Kline) "the founder of abstract algebra."

Exercises 3.3

Throughout these exercises R denotes a ring unless otherwise specified.

1. In each case decide whether A is an ideal of the ring R . Support your answer.
 - (a) $R = \mathbb{C}$, $A = \mathbb{Z}$
 - (b) $R = \mathbb{Z} \times \mathbb{Z}$, $A = \{(k, k) \mid k \in \mathbb{Z}\}$
 - (c) $R = \begin{bmatrix} \mathbb{R} & \mathbb{R} \\ 0 & \mathbb{R} \end{bmatrix}$, $A = \begin{bmatrix} 0 & \mathbb{R} \\ 0 & \mathbb{R} \end{bmatrix}$
 - (d) $R = \begin{bmatrix} \mathbb{Z} & \mathbb{Z} \\ 0 & \mathbb{Z} \end{bmatrix}$, $A = \begin{bmatrix} \mathbb{Z} & 2\mathbb{Z} \\ 0 & \mathbb{Z} \end{bmatrix}$
 - (e) $R = \begin{bmatrix} \mathbb{R} & \mathbb{R} \\ 0 & \mathbb{R} \end{bmatrix}$, $A = \begin{bmatrix} \mathbb{Z} & \mathbb{R} \\ 0 & \mathbb{Z} \end{bmatrix}$
 - (f) $R = \mathbb{Z}(i)$, $A = \{n + ni \mid n \in \mathbb{Z}\}$
2. If $R = \begin{bmatrix} S & S \\ 0 & S \end{bmatrix}$ and $A = \begin{bmatrix} 0 & S \\ 0 & 0 \end{bmatrix}$, S any ring, show that A is an ideal of R and describe the cosets in R/A .
3. If A is an ideal of R , complete the proof of Theorem 1 by verifying that
 - (a) $1 + A$ is the unity of R/A .
 - (b) The associative and distributive laws hold in R/A .
 - (c) If R is commutative, so also is R/A .
4. (a) If m is an integer, show that $mR = \{mr \mid r \in R\}$ and $A_m = \{r \in R \mid mr = 0\}$ are ideals of R .

 (b) If $R = \mathbb{Z}_n$, show that every ideal of R has the form mR for some $m \in \mathbb{Z}$.
5. (a) If A is an ideal of R and B is an ideal of S , show that $A \times B$ is an ideal of $R \times S$.

 (b) Show that every ideal \mathcal{A} of $R \times S$ has the form $\mathcal{A} = A \times B$ as in (a). [Hint: $A = \{a \in R \mid (a, 0) \in \mathcal{A}\}$.]

 (c) Show that the maximal ideals of $R \times S$ are either of the form $A \times S$ where A is maximal in R , or of the form $R \times B$, B maximal in S .
6. If A is an ideal of R , show that $M_2(A)$ is an ideal of $M_2(R)$.
7. Show that $\mathbb{Z} \times 0$ and $0 \times \mathbb{Z}$ are prime ideals of $\mathbb{Z} \times \mathbb{Z}$.
8. If A and B are ideals of R such that $A \cap B = 0$, show that $ab = 0 = ba$ for all $a \in A$ and $b \in B$.
9. Let $R = \mathbb{Z}(i)$ be the ring of gaussian integers. In each case find the number of elements in the factor ring R/A and describe the cosets.
 - (a) $A = Ri$
 - (b) $A = R(1 - i)$
 - (c) $A = R(1 + 2i)$
 - (d) $A = R(1 + 3i)$

[Hint: $(1 + 2i)(1 - i) = 3 + i$ and $(1 + 3i)(1 - 3i) = 10$.]
10. If R is a simple ring, show that $Z(R)$ is a field. Show that the converse is not true by considering $R = \begin{bmatrix} F & F \\ 0 & F \end{bmatrix}$ where F is a field.
11. (a) If R is a simple ring and $n \in \mathbb{Z}$, show that either $nR = 0$, or $nr = 0$, $r \in R$, implies that $r = 0$.

 (b) Conclude that R has characteristic 0 or a prime.

12. If $X \subseteq R$ is a nonempty subset of a commutative ring R , define the **annihilator** of X by $\text{ann}(X) = \{a \in R \mid ax = 0 \text{ for all } x \in X\}$.
- Show that $\text{ann}(X)$ is an ideal of R .
 - If $X \subseteq Y$, show that $\text{ann}(Y) \subseteq \text{ann}(X)$.
 - Show that $\text{ann}(X \cup Y) = \text{ann}(X) \cap \text{ann}(Y)$.
 - Show that $X \subseteq \text{ann}[\text{ann}(X)]$.
 - Show that $\text{ann}(X) = \text{ann}\{\text{ann}(X)\}$.
13. Give an example where R/A is commutative but R is not.
14. If X and Y are additive subgroups of R , define $X + Y = \{x + y \mid x \in X, y \in Y\}$.
- Show that $X + Y$ is an additive subgroup that contains both X and Y .
 - If A and B are ideals of R , show that $A + B$ is an ideal of R .
 - If A is an ideal of R and S is a subring of R , show that $A + S$ is a subring of R .
15. If A is an ideal of R , show that $A \cap S$ is an ideal of S for all subrings S of R .
16. If A is an ideal of R , show that R/A is commutative if and only if $rs - sr \in A$ for all $r, s \in R$.
17. Let $Z = Z(R)$ denote the center of a ring R .
- When is Z an ideal of R ? Justify your answer.
 - If R is simple, show that Z is a field.
 - If R/Z is cyclic as an additive group, show that R is commutative. (This is the analogue for rings of Theorem 2 §2.9.)
18. Let A, B and C be ideals of a ring R .
- Show that $A \cap B$ and $A + B = \{a + b \mid a \in A, b \in B\}$ are ideals of R .
 - If $A \subseteq B$ and $A \subseteq C$, show that $\frac{B}{A} \cap \frac{C}{A} = \frac{B \cap C}{A}$ and $\frac{B}{A} + \frac{C}{A} = \frac{B+C}{A}$.
19. If A is an ideal of R , show that R/A has no nonzero nilpotents if and only if $r^2 \in A$ implies $r \in A$. [Hint: Exercise 11 §3.1.]
20. Let R be a commutative ring.
- Show that every maximal ideal of R is prime.
 - If R is finite, show that every prime ideal is maximal.
 - Is every prime ideal of \mathbb{Z} maximal? Justify your answer.
21. In each case show that, if R has the given property, so does any factor ring R/A .
- Boolean ($r^2 = r$ for all $r \in R$).
 - Regular (for all $r \in R$, $rsr = r$ for some $s \in R$).
 - Every element is a unit or a nilpotent.
22. Let A be an ideal of R consisting of nilpotent elements.
- If R/A has no idempotents except 0 and 1, show that R has the same property.
 - If $u + A$ is a unit in R/A , show that u is a unit in R .
 - Show R/A has no units except 1 if and only if $R^* = 1 + A$.
23. In each case find all maximal ideals of R .
- $R = \mathbb{Z}_5$
 - $R = \mathbb{Z}_8$
 - $R = \mathbb{Z}_{10}$
24. An additive subgroup L of R is called a **left ideal** if $Ra \subseteq L$ for all $a \in L$. Show that R is a division ring if and only if 0 and R are the only left ideals of R (extending Theorem 5). [Hint: Ra is a left ideal for each $a \in R$.]
25. Let R be a commutative ring. Write $a|b$ if $b = ra$ for some $r \in R$.
- Show that $Rab \subseteq Ra \cap Rb$ for all $a, b \in R$.
 - If $Ra + Rb = R$ (see Exercise 14), show that $Rab = Ra \cap Rb$.
 - Show that $u \in R$ is a unit if and only if $Ru = R$.

- (d) Show that Rp is a prime ideal if and only if $p|ab$ implies that $p|a$ or $p|b$.
- (e) If R is an integral domain, show that $Ra = Rb$ if and only if $a = ub$ for some unit $u \in R$.
26. Let A, B , and C be ideals of R and define

$$AB = \{a_1b_1 + a_2b_2 + \cdots + a_nb_n \mid a_i \in A, b_i \in B, n \geq 1\}.$$
- (a) Show that AB is an ideal of R and $AB \subseteq A \cap B$.
- (b) Show that $A(B + C) = AB + AC$ and $(B + C)A = BA + CA$. (Exercise 14.)
- (c) Show that $AR = A = RA$.
- (d) Show that $A(BC) = (AB)C$.
27. If $a \in R$, write $RaR = \{r_1as_1 + r_2as_2 + \cdots + r_nas_n \mid r_i, s_i \in R, n \geq 1\}$. Show that RaR is an ideal of R containing a , and that it is contained in any such ideal.
28. If $e^2 = e \in R$ and A is an ideal of R , show that $eAe = eRe \cap A$, that this is an ideal of eRe , and that every ideal of eRe occurs in this way.
29. Let $R = \begin{bmatrix} F & F \\ 0 & F \end{bmatrix}$, F a field, show that $0, R, \begin{bmatrix} 0 & F \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} F & F \\ 0 & 0 \end{bmatrix}$, and $\begin{bmatrix} 0 & F \\ 0 & F \end{bmatrix}$ are the only ideals of R .
30. If X is an ideal of $\mathbb{H}(R)$ and $2 \in R^*$, show that $X = \mathbb{H}(A)$, where $A = X \cap R$ is an ideal of R . (See Exercise 32 §3.2.)
31. Show that $\mathbb{Z}_2(i)$ has a unique proper ideal $A \neq 0$.
32. (a) Show that $\mathbb{Z}_3(\sqrt{2})$ is a field.
(b) Show that $\mathbb{Z}_2(\sqrt{2})$ has a unique proper ideal $A \neq 0$.
33. If R is commutative, let $N(R) = \{a \in R \mid a \text{ is nilpotent}\}$ —the **nil radical** of R .
(a) Show that $N(R)$ is an ideal of R . [Hint: Theorem 4 §3.1.]
(b) Show that $N[R/N(R)] = 0$.
(c) Show that $N(R)$ need not be an ideal if R is not commutative.
(d) Show that $N(R)$ is contained in the intersection of all prime ideals of R . (In fact, this is equality by Example 4 in Appendix C.)
34. A ring R is called a **local ring** if the set $J(R)$ of nonunits in R forms an ideal.
(a) Show that every division ring R is local. Describe $J(R)$.
(b) If p is a prime, show that $\mathbb{Z}_{(p)} = \{\frac{n}{m} \in \mathbb{Q} \mid p \text{ does not divide } m\}$ is local. Describe $J(\mathbb{Z}_{(p)})$.
(c) If p is a prime and $n \geq 1$, show that \mathbb{Z}_{p^n} is local. Describe $J(\mathbb{Z}_{p^n})$.
(d) If R is local, show that $R/J(R)$ is a division ring.
(e) Let R be local and let $A \subseteq J(R)$ be an ideal of R . Show that R/A is local and $J(R/A) = \{r + A \mid r \in J(R)\}$.
35. Let R be an integral domain and regard $R \subseteq Q$, where Q is the field of quotients (Theorem 5 §3.2). If P is a prime ideal of R , write $M = R \setminus P = \{u \in R \mid u \notin P\}$.
(a) Show that $1 \in M$ and M is closed under multiplication.
(b) Show that $R_P = \{r/u \mid r \in R, u \in M\}$ is a subring of Q .
(c) Show that R_P is a local ring (Exercise 34) called the **localization** of R at P .
36. Let A be an ideal of a ring R consisting of nilpotent elements and assume that R/A is a division ring.
(a) Show that R is local (Exercise 34) and $R^* = R \setminus A$. [Hint: Example 17 §3.1.]
(b) Show that $(1 + A) \triangleleft R^*$ and $R^*/(1 + A) \cong (R/A)^*$ as groups.
(c) Assume that R is commutative and $n \in R^*$ for all $n \geq 2$. Show that $(A, +) \cong 1 + A$ as groups. [Hint: $a \mapsto e^a$; see the discussion following Example 17 §3.1.]

3.4 HOMOMORPHISMS

A ring R is a set with the structure of an additive abelian group and a multiplicative monoid, together with the distributive laws. In this section, we are interested in the structure-preserving mappings $\theta : R \rightarrow S$, where S is another ring. In Section 2.10, the structure-preserving mappings from one group to another (the homomorphisms) turned out to be just those that preserved the operation. A ring has two operations, which suggests that $\theta : R \rightarrow S$ is structure-preserving if it preserves both addition and multiplication. However, in a ring R , the unity 1_R is also part of the structure, so we require θ to preserve the unity: $\theta(1_R) = 1_S$. This requirement is automatic for groups but it can fail in general for rings (Example 6 below).

If R and S are rings, a mapping $\theta : R \rightarrow S$ is called a **ring homomorphism** if, for all r and r_1 in R :

- | | |
|--|-----------------------------------|
| (1) $\theta(r + r_1) = \theta(r) + \theta(r_1)$. | θ preserves addition |
| (2) $\theta(rr_1) = \theta(r) \cdot \theta(r_1)$. | θ preserves multiplication |
| (3) $\theta(1_R) = 1_S$. | θ preserves the unity |

If R and S are general rings, the mapping θ is called a **general ring homomorphism** if (1) and (2) hold, but possibly not (3).

Example 1. If A is an ideal of R , the coset map $R \rightarrow R/A$ given by $r \mapsto r + A$ is an onto ring homomorphism.

Example 2. The mapping $k \mapsto \bar{k}$ from \mathbb{Z} to \mathbb{Z}_n is an onto ring homomorphism.

Example 3. If R_1 and R_2 are rings, the projections $\pi_1 : R_1 \times R_2 \rightarrow R_1$ and $\pi_2 : R_1 \times R_2 \rightarrow R_2$ are onto ring homomorphisms, where $\pi_1(r_1, r_2) = r_1$ and $\pi_2(r_1, r_2) = r_2$.

Example 4. If $\theta : \begin{bmatrix} R & R \\ 0 & R \end{bmatrix} \rightarrow R \times R$ is given by $\theta \begin{bmatrix} r & s \\ 0 & t \end{bmatrix} = (r, t)$, show that θ is an onto ring homomorphism.

Solution. The reader should verify that θ is an onto homomorphism of additive groups. We have $\theta \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = (1, 1)$, so θ preserves the unity. Finally

$$\begin{aligned} \theta \left\{ \begin{bmatrix} r & s \\ 0 & t \end{bmatrix} \begin{bmatrix} r' & s' \\ 0 & t' \end{bmatrix} \right\} &= \theta \begin{bmatrix} rr' & rs' + st' \\ 0 & tt' \end{bmatrix} = (rr', tt') = (r, t) \cdot (r', t') \\ &= \theta \begin{bmatrix} r & s \\ 0 & t \end{bmatrix} \cdot \theta \begin{bmatrix} r' & s' \\ 0 & t' \end{bmatrix}. \end{aligned}$$

Hence, θ preserves multiplication. □

Example 5. If R and S are rings, let $\theta : R \rightarrow S$ is an *onto* mapping that preserves addition and multiplication. Show that θ is a ring homomorphism, that is $\theta(1_R) = 1_S$.

Solution. The argument preceding Example 19 §3.1 goes through. □

Example 6. The mapping $\theta : R \rightarrow R \times R$, where $\theta(r) = (r, 0)$ is a (one-to-one) general ring homomorphism that does not preserve the unity if $R \neq 0$.

Ring homomorphisms are homomorphisms of additive abelian groups, which gives the first three preservation properties in the next result. We leave the proofs of the last two as Exercise 10.

Theorem 1. Let $\theta : R \rightarrow R_1$ be a ring homomorphism and let $r \in R$.

- | | |
|---|---|
| (1) $\theta(0) = 0$. | θ preserves zero |
| (2) $\theta(-r) = -\theta(r)$. | θ preserves negatives |
| (3) $\theta(kr) = k\theta(r)$ for all $k \in \mathbb{Z}$. | θ preserves \mathbb{Z} -multiplication |
| (4) $\theta(r^n) = \theta(r)^n$ for all $n \geq 0$ in \mathbb{Z} | } |
| (5) If $u \in R^*$, $\theta(u^k) = \theta(u)^k$ for all $k \in \mathbb{Z}$ | θ preserves powers |

By a **rational expression** in a ring R we mean a formula made up of letters representing elements of R that are combined using addition, subtraction, multiplication, division (by units), and multiplication by integers. Thus, $r^2su^5 - 3su^{-2}r + 2$ is a rational expression where, of course, u is a unit in R and 2 means $2 \cdot 1_R$. Because of Theorem 1 (and the ring axioms), a ring homomorphism $\theta : R \rightarrow S$ preserves rational expressions. For example, if we write $\theta(x) = \bar{x}$ for every $x \in R$, then

$$\theta(r^2su^5 - 3su^{-2}r + 2) = \bar{r}^2\bar{s}\bar{r}^5 - 3\bar{s}\bar{u}^{-2}\bar{r} + \bar{2}.$$

In particular, if $r \in R$ is a unit, an idempotent, or a nilpotent, the same is true of the element $\bar{r} = \theta(r)$ in R_1 .

The fact that ring homomorphisms preserve rational expressions is very useful. One reason is that, in many rings derived from a ring R (for example $M_n(R)$), we define the operations using rational expressions from R . Hence, a ring homomorphism $R \rightarrow S$ often induces a homomorphism of the derived ring in a natural way. Here is an example.

Example 7. If $\theta : R \rightarrow S$ is a ring homomorphism, show that $\bar{\theta} : M_2(R) \rightarrow M_2(S)$ is also a ring homomorphism where

$$\bar{\theta} \begin{bmatrix} r & s \\ t & u \end{bmatrix} = \begin{bmatrix} \theta(r) & \theta(s) \\ \theta(t) & \theta(u) \end{bmatrix}, \quad \text{for all } \begin{bmatrix} r & s \\ t & u \end{bmatrix} \text{ in } M_2(R).$$

Solution. We leave to the reader the verification that $\bar{\theta}$ preserves addition and the unity. For convenience, write $\theta(r) = \bar{r}$ for all $r \in R$. Then

$$\begin{aligned} \bar{\theta} \left\{ \begin{bmatrix} r & s \\ t & u \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right\} &= \begin{bmatrix} \overline{ra+sc} & \overline{rb+sd} \\ \overline{ta+uc} & \overline{tb+ud} \end{bmatrix} \\ &= \begin{bmatrix} \bar{r}\bar{a} + \bar{s}\bar{c} & \bar{r}\bar{b} + \bar{s}\bar{d} \\ \bar{t}\bar{a} + \bar{u}\bar{c} & \bar{t}\bar{b} + \bar{u}\bar{d} \end{bmatrix} \\ &= \begin{bmatrix} \bar{r} & \bar{s} \\ \bar{t} & \bar{u} \end{bmatrix} \begin{bmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{bmatrix} \\ &= \bar{\theta} \begin{bmatrix} r & s \\ t & u \end{bmatrix} \cdot \bar{\theta} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \end{aligned}$$

Hence, $\bar{\theta}$ preserves multiplication, and so is a ring homomorphism. \square

Another way in which the preservation of rational expressions by homomorphisms is useful is in showing that an equation in a ring R has no solution in R . The reason is that, if $\theta : R \rightarrow S$ is a homomorphism and if an equation has a

solution in R , then (because θ preserves the whole equation) it has a solution in S . Thus by showing that no solution exists in S , we show that no solution can exist in R . This approach is useful because the ring S is often much simpler than R , so the task of showing that no solution exists is easier. We give two examples.

Example 8. Show that $x^3 - 5x^2 - x - 17 = 0$ has no solution in \mathbb{Z} .

Solution. Consider the homomorphism $\theta : \mathbb{Z} \rightarrow \mathbb{Z}_5$ given by $\theta(k) = \bar{k}$. Suppose that $n \in \mathbb{Z}$ is a solution: $n^3 - 5n^2 - n - 17 = 0$. Applying θ gives $\bar{n}^3 - 5\bar{n}^2 - \bar{n} - \bar{17} = \bar{0}$ in \mathbb{Z}_5 ; that is, $\bar{n}^3 - \bar{n} - \bar{2} = \bar{0}$. But \bar{n} is one of $\bar{0}, \bar{1}, \bar{2}, \bar{3}$, or $\bar{4}$ in \mathbb{Z}_5 , and a direct check shows that none of these satisfies the equation $\bar{n}^3 - \bar{n} - \bar{2} = \bar{0}$. Hence, no solution of the original equation could exist in \mathbb{Z} . \square

Example 9. Show that $m^3 - 6n^3 = 3$ has no solution in \mathbb{Z} .

Solution. Our first temptation is to reduce this modulo 6, obtaining $m^3 = 3$ in \mathbb{Z}_6 . But this has a solution ($m = 3$) in \mathbb{Z}_6 , so there is no gain here. However, in \mathbb{Z}_7 the equation becomes $m^3 + n^3 = 3$. But the only cubes in \mathbb{Z}_7 are 0, 1, and 6, and the sum of two of these is one of 0, 1, 2, 5, or 6. Because 3 is not in this list, there is no solution in \mathbb{Z}_7 and hence none in \mathbb{Z} . \square

Our next theorem discusses an important homomorphism of rings of prime characteristic, that will be needed later. The proof depends on a fact about the binomial coefficients $\binom{p}{k} = \frac{p!}{k!(p-k)!}$ that is important in its own right.

Lemma 1. If p is a prime then p divides $\binom{p}{k}$ for each $k = 1, 2, \dots, p-1$.

Proof. The definition of $\binom{p}{k}$ gives $p! = \binom{p}{k} k!(p-k)!$, so p divides the product $\binom{p}{k} k!(p-k)!$. Hence, Euclid's lemma (Theorem 6 §1.2) shows that p must either divide $\binom{p}{k}$ or divide some factor of $k!(p-k)!$. But this latter outcome is impossible because $1 \leq k \leq p-1$, so p must divide $\binom{p}{k}$ as asserted. \blacksquare

Theorem 2. Let $R \neq 0$ be a commutative ring of prime characteristic p , and define

$$\varphi : R \rightarrow R \quad \text{by} \quad \varphi(r) = r^p \text{ for all } r \in R.$$

Then φ is a ring homomorphism (the **Frobenius Endomorphism**). If R is a finite field then φ is an isomorphism (the **Frobenius Automorphism**).

Proof. Clearly $\varphi(1) = 1$, and $\varphi(rs) = \varphi(r)\varphi(s)$ because R is commutative. We have

$$\varphi(r+s) = r^p + \binom{p}{1}r^{p-1}s + \cdots + \binom{p}{p-1}rs^{p-1} + s^p$$

for all $r, s \in R$ by the binomial theorem. But p divides each of the coefficients $\binom{p}{1}, \dots, \binom{p}{p-1}$ by Lemma 1, so each of these coefficients is zero in R because R has characteristic p . Hence, $\varphi(r+s) = \varphi(r) + \varphi(s)$, so φ is a ring homomorphism.

If R is a field, observe that $\ker \varphi$ is an ideal of R (see Theorem 3 below). Moreover, $\ker \varphi \neq R$ because $\varphi(1_R) = 1_R$. Hence, $\ker \varphi = 0$ because R is a simple ring, so φ is one-to-one (being an additive group homomorphism). If R is finite, then φ is also onto (Theorem 2 §0.3), and so φ is an isomorphism. \blacksquare

If Q is an infinite field, a ring homomorphism $Q \rightarrow Q$ can exist that is one-to-one but not onto. If p is a prime, let Q denote the field of quotients of the integral

domain $\mathbb{Z}_p[x]$ (see Chapter 4). Then $\text{char } Q = p$ so the Frobenius endomorphism $q \mapsto q^p$ is a one-to-one ring homomorphism $Q \rightarrow Q$, but it is not onto.

The Isomorphism Theorem

A ring homomorphism $\theta : R \rightarrow S$ is, in particular, a homomorphism of additive groups. Hence, it has a kernel and an image

$$\ker \theta = \{a \in R \mid \theta(a) = 0\} \quad \text{and} \quad \text{im } \theta = \theta(R) = \{\theta(r) \mid r \in R\}.$$

These are additive subgroups of R and S , respectively and, by Theorem 3 §2.10, $\theta : R \rightarrow S$ is one-to-one if and only if $\ker \theta = 0$. We also have the ring theoretic analogue of Theorem 1 §2.10.

Theorem 3. *Let $\theta : R \rightarrow S$ be a ring homomorphism.*

- (1) $\theta(R)$ is a subring of S .
- (2) $\ker \theta$ is an ideal of R .

Proof. (1) We know that $\theta(R)$ is an additive subgroup of S , and it is closed under multiplication because $\theta(a) \cdot \theta(b) = \theta(ab)$. Finally, our insistence that ring homomorphisms preserves the unity gives $1_S = \theta(1_R) \in \theta(R)$.

(2) Group theory shows that $\ker \theta$ is an additive subgroup of R . If $r \in R$ and $a \in \ker \theta$, then $\theta(ra) = \theta(r) \cdot \theta(a) = \theta(r) \cdot 0 = 0$. Thus $ra \in \ker \theta$ and, similarly, $ar \in \ker \theta$. Hence, $\ker \theta$ is an ideal of R . \blacksquare

As for groups, part (2) of Theorem 3 has a converse: Every ideal A of a ring R is the kernel of some ring homomorphism $R \rightarrow S$. In fact, the coset map $\varphi : R \rightarrow R/A$ is a ring homomorphism (in fact onto) with $\ker \varphi = A$.

We now come to the most important theorem of this section, the ring analogue of the isomorphism theorem for groups.

Theorem 4. Isomorphism Theorem. *Let $\theta : R \rightarrow S$ be a ring homomorphism and write $A = \ker \theta$. Then θ induces the ring isomorphism*

$$\bar{\theta} : R/A \rightarrow \theta(R) \quad \text{given by} \quad \bar{\theta}(r+A) = \theta(r) \quad \text{for all } r \in R.$$

Proof. The kernel A of θ is an ideal of R by Theorem 3, so R/A is a ring. Given a and b in R , compute

$$a + A = b + A \iff (a - b) \in A \iff \theta(a - b) = 0 \iff \theta(a) = \theta(b).$$

This shows that $\bar{\theta}$ is well defined and one-to-one. Because $\bar{\theta}$ is clearly onto $\theta(R)$, it remains to show that $\bar{\theta}$ is a ring homomorphism. Now

$$\bar{\theta}(1_{R/A}) = \bar{\theta}(1_R + A) = \theta(1_R) = 1_S$$

is the unity of $\theta(R)$, so $\bar{\theta}$ preserves the unity. For $a, b \in R$, we have

$$\bar{\theta}[(a + A)(b + A)] = \bar{\theta}(ab + A) = \theta(ab) = \theta(a) \cdot \theta(b) = \bar{\theta}(a + A) \cdot \bar{\theta}(b + A),$$

so $\bar{\theta}$ preserves multiplication. Similarly, $\bar{\theta}$ preserves addition and so is a ring isomorphism. \blacksquare

As for groups, the ring isomorphism theorem is very useful and reveals structure whenever it is used. We devote much of the remainder of this section to illustrations of how it is employed. We begin with three examples (Examples 10–12) involving specific rings. The general theme is: To show that A is an ideal of R and $R/A \cong S$, find an onto ring homomorphism $\theta : R \rightarrow S$ with $\ker \theta = A$.

Example 10. Let A and B be ideals of R and S , respectively. Show that $A \times B$ is an ideal of $R \times S$, and $\frac{R \times S}{A \times B} \cong \frac{R}{A} \times \frac{S}{B}$.

Solution. Define $\theta : R \times S \rightarrow \frac{R}{A} \times \frac{S}{B}$ by $\theta(r, s) = (r + A, s + B)$. Then θ is an onto ring homomorphism and $\ker \theta = A \times B$, so the isomorphism theorem does it. \square

It is worth noting that *every* ideal of $R \times S$ has the form $A \times B$, where A and B are ideals of R and S , respectively (Exercise 5(b) §3.3). Hence, Example 10 describes all homomorphic images of $R \times S$.

Similarly, every ideal of $M_n(R)$ has the form $M_n(A)$ for some ideal A of R , (Lemma 3 §3.3) so the next example describes all homomorphic images of $M_n(R)$.

Example 11. If A is an ideal of a ring R , show that $M_n(A)$ is an ideal of $M_n(R)$, and that $\frac{M_n(R)}{M_n(A)} \cong M_n\left(\frac{R}{A}\right)$.

Solution. If $r \in R$, we write $\bar{r} = r + A$ in R/A for convenience. Then the coset map $\varphi : R \rightarrow R/A$, given by $\varphi(r) = \bar{r}$ for all $r \in R$, is an onto ring homomorphism. Hence, φ induces the homomorphism

$$\bar{\varphi} : M_n(R) \rightarrow M_n\left(\frac{R}{A}\right) \quad \text{given by } \bar{\varphi}[a_{ij}] = [\bar{a}_{ij}] \text{ for all } [a_{ij}] \in M_n(R).$$

Since φ is a ring homomorphism, it is a routine verification that the same is true of $\bar{\varphi}$ (Example 7 is the case $n = 2$). Moreover $\ker \bar{\varphi} = M_n(A)$ because $\ker \varphi = A$. Now the isomorphism theorem applies. \square

Example 12. If $m|n$, find an ideal A of \mathbb{Z}_n such that $\mathbb{Z}_n/A \cong \mathbb{Z}_m$.

Solution. This can be solved directly by examining the factor rings of \mathbb{Z}_n , but (as is often the case) it is easier to let the isomorphism theorem do the work. Because $\mathbb{Z}_n = \{k + n\mathbb{Z} \mid k \in \mathbb{Z}\}$, there is a natural map $\theta : \mathbb{Z}_n \rightarrow \mathbb{Z}_m$ given by $\theta(k + n\mathbb{Z}) = k + m\mathbb{Z}$. This mapping is well defined because $m|n$:

$$k + n\mathbb{Z} = k' + n\mathbb{Z} \Rightarrow n|(k - k') \Rightarrow m|(k - k') \Rightarrow k + m\mathbb{Z} = k' + m\mathbb{Z}.$$

With this θ is clearly an onto ring homomorphism, so we are done with $A = \ker \theta$ by the isomorphism theorem. In fact,

$$\ker \theta = \{k + n\mathbb{Z} \mid k + m\mathbb{Z} = m\mathbb{Z}\} = \{mq + n\mathbb{Z} \mid q \in \mathbb{Z}\} = \{m\bar{q} \mid q \in \mathbb{Z}\} = m\mathbb{Z}_n. \quad \square$$

Theorem 5. If R is any ring, then $\mathbb{Z}1_R = \{k1_R \mid k \in \mathbb{Z}\}$ is a subring of R that is contained in the center of R . Moreover,

- (1) If R has characteristic $n > 0$, then $\mathbb{Z}1_R \cong \mathbb{Z}_n$.
- (2) If R has characteristic 0, then $\mathbb{Z}1_R \cong \mathbb{Z}$.

Proof. Define $\theta : \mathbb{Z} \rightarrow R$ by $\theta(k) = k1_R$ for all $k \in \mathbb{Z}$. This map is a ring homomorphism by Theorem 2 §3.1, so $\mathbb{Z}1_R = \theta(\mathbb{Z})$ is a subring of R by Theorem 3. Moreover, $\mathbb{Z}1_R$ is contained in the center of R because $r(k1_R) = kr = (k1_R)r$ for all $r \in R$ and $k \in \mathbb{Z}$ by the distributive laws and Theorem 2 §3.1 (verify).

We have $\ker \theta = \{k \in \mathbb{Z} \mid k1_R = 0\}$. If R has characteristic $n > 0$, then we have $\ker \theta = n\mathbb{Z}$ by Theorem 3 §3.1. Hence, $\mathbb{Z}1_R = \theta(\mathbb{Z}) \cong \mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$ by the isomorphism theorem, proving (1). If R has characteristic 0, then $\ker \theta = 0$ and (2) again follows by the isomorphism theorem. ■

Theorem 5 is particularly important if R is a field. In this case a subring S of R is called a **subfield** of R if it is itself a field, or equivalently if $s^{-1} \in S$ whenever $0 \neq s \in S$. Now the characteristic of a field R is either 0 or a prime p . If $\text{char } R = p$, Theorem 5 shows that R contains a central subfield $\mathbb{Z}1_R \cong \mathbb{Z}_p$.

If $\text{char } R = 0$, the central subring $\mathbb{Z}1_R$ is isomorphic to \mathbb{Z} . In this case define

$$Q = \{uv^{-1} \mid u, v \text{ in } \mathbb{Z}1_R, v \neq 0\}.$$

This is easily verified to be a central subfield of R , and we claim that $Q \cong \mathbb{Q}$. Indeed, the map $\varphi : \mathbb{Q} \rightarrow Q$ given by $\varphi(n/m) = (n1_R)(m1_R)^{-1}$ is a ring isomorphism. We leave the verification to the reader with the observation that the proof that φ is well defined and one-to-one uses the following fact: Since $\text{char } R = 0$, if $n \in \mathbb{Z}$ then $n1_R = 0$ if and only if $n = 0$. This proves the

Corollary. Every field R contains a central subfield isomorphic to \mathbb{Z}_p or \mathbb{Q} according as $\text{char } R = p$ or $\text{char } R = 0$.

Because of this result, the fields \mathbb{Z}_p and \mathbb{Q} are called **prime fields**. They are important in field theory and we mention them again in Chapter 6.

We can reduce many questions about general rings (with no unity) to the case of rings by a standard construction. If R is a general ring, consider the set

$$R^1 = \mathbb{Z} \times R$$

and define operations on R^1 as follows:

$$\begin{aligned}(n, r) + (m, s) &= (n + m, r + s) \\ (n, r)(m, s) &= (nm, ns + mr + rs)\end{aligned}$$

Then R^1 is a ring with unity $(1, 0)$ as the reader can easily verify, and the mapping $\theta : R^1 \rightarrow \mathbb{Z}$ defined by $\theta(n, r) = n$ is an onto ring homomorphism for which $\ker \theta = \{(0, r) \mid r \in R\}$. The mapping $\sigma : R \rightarrow \ker \theta$ with $\sigma(r) = (0, r)$ is a one-to-one, onto general ring homomorphism (preserves addition and multiplication). Hence, we may regard R as a subset of R^1 by identifying $r = (0, r)$ for all $r \in R$. This being done, R is an *ideal* of R^1 and the isomorphism theorem gives

Theorem 6. If R is a general ring, a ring R^1 exists, containing R as an ideal such that $R^1/R \cong \mathbb{Z}$.

Decompositions of Rings

When trying to ascertain the structure of a ring, it is useful to have a condition that ensures that a ring R is isomorphic to a direct product of two subrings. We need the following notion: If A and B are ideals of a ring R , define their **sum** $A + B$ by

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

It is not difficult to show that $A + B$ is again an ideal of R , the smallest that contains both A and B . Our interest is in the case when $A + B = R$.

Theorem 7. Let R be a ring with ideals A and B such that

$$R = A + B \quad \text{and} \quad A \cap B = \{0\}.$$

Let $1 = e + f$ in R where $e \in A$ and $f \in B$. Then

- (1) A and B are rings with unities e and f respectively (both central in R).
- (2) $R \cong A \times B$ as rings.

Proof. (1) If $a \in A$ then $a = ae + af$ so $a - ae = af \in A \cap B = 0$. It follows that $a = ae$; and similarly $a = ea$. Hence, e is the unity for A (and so $e^2 = e$). In the same way, $f = f^2$ is the unity of B . They are central in R by Example 5 §3.3.

(2) Define $\theta : A \times B \rightarrow R$ by $\theta(a, b) = a + b$. Then θ is onto because $R = A + B$, $\theta(e, f) = 1$, and θ is easily verified to be a homomorphism of additive groups. Moreover, if $\theta(a, b) = 0$ then $a = -b \in A \cap B = 0$, and it follows that θ is one-to-one. To see that θ preserves multiplication, note that if $a \in A$ and $b \in B$ then $ab \in A \cap B = 0$, and similarly $ba = 0$. But then, if $a' \in A$ and $b' \in B$:

$$\theta(a, b) \cdot \theta(a', b') = (a + b)(a' + b') = aa' + bb' = \theta(aa', bb') = \theta[(a, b)(a', b')].$$

Hence, θ is a ring homomorphism, and so is an isomorphism. ■

There is a converse to part (1) in Theorem 7. If $e^2 = e \in R$ is central, then $A = eR = Re = eRe$ is an ideal. Moreover, $f = 1 - e$ is also central, and $B = fRf$ is also an ideal. One verifies that $R = A + B$ and $A \cap B = 0$, so Theorem 7 gives

Corollary. If $e^2 = e \in R$ is central, then $R \cong eRe \times (1 - e)R(1 - e)$.

Let A and B be ideals of a ring R . Theorem 7 characterizes when R is isomorphic to $A \times B$. Part (2) of the next theorem gives essentially the same result in the form that R is isomorphic to $(R/A) \times (R/B)$.

Theorem 8. Chinese Remainder Theorem.⁵⁶ Let A and B be ideals of R .

- (1) If $A + B = R$ then $\frac{R}{A \cap B} \cong \frac{R}{A} \times \frac{R}{B}$.
- (2) If $A + B = R$ and $A \cap B = 0$ then $R \cong \frac{R}{A} \times \frac{R}{B}$.

Proof. Since (1) implies (2) because $\frac{R}{0} \cong R$, we need only prove (1). Define

$$\psi : R \rightarrow \frac{R}{A} \times \frac{R}{B} \quad \text{by} \quad \psi(r) = (r + A, r + B) \text{ for all } r \in R.$$

Then ψ is a ring homomorphism and $\ker \psi = A \cap B$. Hence, by the isomorphism theorem, it remains to show that ψ is onto. Since $A + B = R$, write $1 = a + b$ where $a \in A$ and $b \in B$. Given $(s + A, t + B) \in \frac{R}{A} \times \frac{R}{B}$ where s and t are in R , let $r = sb + ta$. Then

$$s - r = s(1 - b) - ta = (s - t)a \in A$$

so $s + A = r + A$. Similarly $t + B = r + B$, and so $\psi(r) = (s + A, t + B)$. This shows that ψ is onto, as required. ■

⁵⁶The name derives from the fact that a special case ($R = \mathbb{Z}$) of the theorem was known to the Chinese in the first century AD.

Corollary 1. If m and n are relatively prime, then $\mathbb{Z}_{mn} \cong \mathbb{Z}_m \times \mathbb{Z}_n$.

Proof. We are asking that $\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ so, taking $R = \mathbb{Z}$, $A = m\mathbb{Z}$, and $B = n\mathbb{Z}$ in Theorem 8, we must prove that $m\mathbb{Z} + n\mathbb{Z} = \mathbb{Z}$ and $m\mathbb{Z} \cap n\mathbb{Z} = mn\mathbb{Z}$. The first follows from $\gcd(m, n) = 1$ since then $1 = mp + nq$ where $p, q \in \mathbb{Z}$, so 1 is in the ideal $m\mathbb{Z} + n\mathbb{Z}$. Hence, $m\mathbb{Z} + n\mathbb{Z} = \mathbb{Z}$ by Theorem 3 §3.3.

If $k \in m\mathbb{Z} \cap n\mathbb{Z}$, then $m|k$ and $n|k$, so $mn|k$, again because $\gcd(m, n) = 1$ (Theorem 5 §1.2). Hence, $m\mathbb{Z} \cap n\mathbb{Z} \subseteq mn\mathbb{Z}$; the other inclusion always holds. ■

Corollary 1 has a useful application to number theory. Recall that we defined the Euler function φ (at the end of Section 2.6) by taking $\varphi(n)$ to be the number of integers in the set $\{1, 2, \dots, n-1\}$ that are relatively prime to n . Hence $\varphi(n) = |\mathbb{Z}_n^*|$, and it is here that Corollary 1 comes into play.

Corollary 2. If $m \geq 2$ and $n \geq 2$ are relatively prime, then $\varphi(mn) = \varphi(m) \cdot \varphi(n)$.

Proof. We have $\mathbb{Z}_{mn}^* \cong (\mathbb{Z}_m \times \mathbb{Z}_n)^* = \mathbb{Z}_m^* \times \mathbb{Z}_n^*$ from Corollary 1, and the result follows because $\varphi(k) = |\mathbb{Z}_k^*|$ for all $k \geq 2$. ■

Emmy Noether (1882–1935) Herman Weyl has described Emmy Noether as “a great mathematician, the greatest, I firmly believe, that her sex has ever produced, and a great woman.” She was born in Bavaria, the daughter of a well-known algebraist Max Noether. She completed her doctorate at Erlangen in 1907 and, in 1916, went to Göttingen to work with David Hilbert. Göttingen was then one of the leading centers of mathematics and, by 1930, Noether had established a fertile and influential research program that was recognized as the primary center of algebraic thought in the world. But, even with the enthusiastic support of Hilbert, she never attained more than an honorary professorship at Göttingen in part because she was a woman. With the rise of Hitler, she was forced to leave because she was a Jew, and she spent the last 2 years of her life at Bryn Mawr college in Pennsylvania.

Her work touched several fields (general relativity and the calculus of variations, among others), but her genius flowered in algebra. However, she published comparatively little (she was most generous in sharing her ideas with others, especially her students). Even so, she created a whole new trend in algebra, emphasizing axiomatic concepts of great generality. To quote the Russian mathematician P.S. Alexandroff, “Emmy Noether taught us to think in a simpler and more general way; in terms of homomorphisms, of ideals—not in terms of complicated algebraic calculations. She therefore opened a path to the discovery of algebraic regularities where previously they had been obscured by complicated specific conditions.” Her 1921 paper on ideal theory was a landmark and has had a profound influence on ring theory and on algebra generally. It emphasized the fundamental importance of certain finiteness conditions, some of which can be traced back to Dedekind. As a result, rings satisfying the so-called ascending chain condition on ideals are now called *noetherian* rings.

Exercises 3.4

Throughout these exercises R denotes a ring unless otherwise specified.

- In each case determine whether the map θ is a ring homomorphism. Support your answer.
 - $\theta : \mathbb{Z}_3 \rightarrow \mathbb{Z}_{12}$, where $\theta(r) = 4r$

- (b) $\theta : \mathbb{Z}_4 \rightarrow \mathbb{Z}_{12}$, where $\theta(r) = 3r$
 (c) $\theta : R \times R \rightarrow R$, where $\theta(r, s) = r + s$
 (d) $\theta : R \times R \rightarrow R$, where $\theta(r, s) = rs$
 (e) $\theta : F(\mathbb{R}, \mathbb{R}) \rightarrow \mathbb{R}$, where $\theta(f) = f(1)$
2. Let $\theta : R \rightarrow S$ be a general ring homomorphism, where R and S are rings. Show that θ is a ring homomorphism if: (a) θ is onto; (b) S is a domain and $\theta(1) \neq 0$.
3. Show that a general ring homomorphism $\theta : \mathbb{Z} \rightarrow \mathbb{Z}$ is either a ring isomorphism or $\theta(k) = 0$ for all $k \in \mathbb{Z}$.
4. Determine all onto ring (general ring) homomorphisms $\mathbb{Z}_{12} \rightarrow \mathbb{Z}_6$.
5. If $\theta : R \rightarrow R_1$ is an onto ring homomorphism, show that $\theta[Z(R)] \subseteq Z(R_1)$. Give an example showing that this need not be equality.
6. If $\theta : R \rightarrow R_1$ is a ring homomorphism and $\text{char } R = n > 0$, show that $\text{char } R_1$ divides n .
7. Show that the composite of two ring homomorphisms is a ring homomorphism.
8. Let R and S be rings and let $\theta : R \rightarrow S$ be a general ring homomorphism (that is, $\theta(1)$ may not be the unity of S). If $\theta(1) = e$, show that $e^2 = e$ in S , $\theta(R) \subseteq eSe$, and $\theta : R \rightarrow eSe$ is a ring homomorphism.
9. Describe the homomorphic images of a division ring.
10. Prove (4) and (5) of Theorem 1.
11. Show that $x^3 - 8x^2 + 5x + 3 = 0$ has no solution $x \in \mathbb{Z}$.
12. Show that $m^3 + 14n^3 = 12$ has no solution in \mathbb{Z} .
13. Show that $7m^2 + 11n^2 = 9$ has no solution in \mathbb{Z} .
14. Show that $n^3 + (n+1)^3 + (n+2)^3 = k^2 + 1$ has no solution in \mathbb{Z} .
15. If $\sigma : R \rightarrow S$ is a ring isomorphism, show that the same is true of the inverse map $\sigma^{-1} : S \rightarrow R$.
16. Show that the set $\text{aut } R$ of all automorphisms of R is a group under composition.
17. Show that the isomorphism relation \cong is an equivalence on the class of all rings.
18. Let $R = \begin{bmatrix} F & F \\ 0 & F \end{bmatrix}$ where F is a field. Determine all homomorphic images of R .
[Hint: Exercise 29 §3.3.]
19. Let $\theta : R \rightarrow S$ be an onto ring homomorphism.
- (a) If A is an ideal of R , show that $\theta(A) = \{\theta(a) \mid a \in A\}$ is an ideal of S .
 - (b) If also $\ker \theta \subseteq A$, show that $\frac{R}{A} \cong \frac{S}{\theta(A)}$.
[Hint: Use the isomorphism theorem where $\alpha : R \rightarrow \frac{S}{\theta(A)}$ is defined by $\alpha(r) = \theta(r) + \theta(A)$ for all $r \in R$.]
20. If $n > 0$ in \mathbb{Z} , describe all the ideals of \mathbb{Z} that contain $n\mathbb{Z}$.
21. Show that there is no ring homomorphism $\mathbb{C} \rightarrow \mathbb{R}$.
22. Let $\theta : R \rightarrow S$ be a ring homomorphism. If $\theta(R)$ and $\ker \theta$ both contain no nonzero nilpotents show that the same is true of R .
23. Let $\theta : R \rightarrow S$ be a ring homomorphism and let $A \subseteq R$ and $B \subseteq S$ be ideals.
- (a) If $\theta(A) \subseteq B$, show that θ induces a unique ring homomorphism $\bar{\theta} : R/A \rightarrow S/B$ such that $\bar{\theta}\varphi = \varphi'\theta$ as shown in the figure (where φ and φ' are the coset maps).
 - (b) Show that (a) applies where R and S are commutative and $A = N(R)$ and $B = N(S)$ are ideals of all nilpotent elements. (See Exercise 33 §3.3.)

$$\begin{array}{ccc}
 R & \xrightarrow{\theta} & S \\
 \varphi \downarrow & & \downarrow \varphi' \\
 R/A & \xrightarrow{\bar{\theta}} & S/B
 \end{array}$$

24. If $u \in R^*$ consider the inner automorphism $\sigma_u : R \rightarrow R$ defined by $\sigma_u(r) = uru^{-1}$ for all $r \in R$ (see Example 18 §3.1). Write $\text{inn } R = \{\sigma_u \mid u \in R^*\}$ for the set of inner automorphisms of R .
- Show that $\text{inn } R$ is a normal subgroup of $\text{aut } R$.
 - If $Z = Z(R)$, show that $Z \cap R^* \triangleleft R^*$ and $R^*/(Z \cap R^*) \cong \text{inn } R$ as groups.
25. If $ab = 1$ in R , write $e = ba$ and define $\sigma : R \rightarrow R$ by $\sigma(r) = bra$.
- Show that $e^2 = e$ and that $\sigma : R \rightarrow eRe$ is a ring isomorphism.
 - Use (a) to show that $ab = 1$ implies that $ba = 1$ if R is a finite ring.
26. If F is a field, find a maximal ideal M of $R = \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in F \right\}$. Describe R/M .
27. Let $R = \begin{bmatrix} S & S \\ 0 & S \end{bmatrix}$ be the upper triangular matrix ring over a ring S . Show that $A = \begin{bmatrix} 0 & S \\ 0 & 0 \end{bmatrix}$ is an ideal of R and $R/A \cong S \times S$.
28. If p is a prime, let $\mathbb{Z}_{(p)} = \{\frac{n}{m} \in \mathbb{Q} \mid p \text{ does not divide } m\}$ and consider the set $J(\mathbb{Z}_{(p)}) = \{\frac{n}{m} \in \mathbb{Z}_{(p)} \mid p \text{ divides } n\}$. Show that $J(\mathbb{Z}_{(p)})$ is an ideal of $\mathbb{Z}_{(p)}$ and $\mathbb{Z}_{(p)}/J(\mathbb{Z}_{(p)}) \cong \mathbb{Z}_p$. (See Exercise 34 §3.3).
29. Consider $R(\omega)$ where $\omega^2 = -1$, as discussed preceding Example 5 §3.2.
- If A is an ideal of R , show that $A(\omega)$ is an ideal of $R(\omega)$ and $\frac{R(\omega)}{A(\omega)} \cong \frac{R}{A}(\omega)$.
 - Show that $3\mathbb{Z}(i)$ is a maximal ideal of $\mathbb{Z}(i)$.
30. If A is an ideal of R , write $\bar{R} = R/A$ and $\bar{r} = r + A$, $r \in R$. If $e^2 = e \in R$, show that $eAe = eRe \cap A$, that this is an ideal of eRe , and that $(eRe)/(eAe) \cong \bar{e}\bar{R}\bar{e}$.
31. Let $R = S \times T$ and write $\bar{S} = \{(s, 0) \mid s \in S\}$. Show that \bar{S} is an ideal of R , $R/\bar{S} \cong T$, and $\bar{S} \cong S$ as rings. What is the unity of \bar{S} ?
32. Prove the **Second Isomorphism Theorem**: If A is an ideal of R and S is a subring of R , then $S + A$ is a subring, A and $S \cap A$ are ideals of $S + A$ and S , respectively, and $(S + A)/A \cong S/(S \cap A)$.
33. Prove the **Third Isomorphism Theorem**: If $A \subseteq B \subseteq R$, where A and B are ideals of R , then $B/A = \{b + A \mid b \in B\}$ is an ideal of R/A and $(R/A)/(B/A) \cong R/B$.
34. Show that every additive subgroup of R is an ideal if and only if $R \cong \mathbb{Z}$ or $R \cong \mathbb{Z}_n$ for some $n \geq 1$.
35. As in the discussion preceding Example 5 §3.2, define $R(\eta)$ to be the set of all formal sums $a + b\eta$, $a, b \in R$, where $\eta^2 = 0$, $a\eta = \eta a$ for all $a \in R$, and $a + b\eta = c + d\eta$ if and only if $a = c$ and $b = d$.
- If A is an ideal of R , show that $A(\eta)$ is an ideal of $R(\eta)$ and $\frac{R(\eta)}{A(\eta)} \cong \frac{R}{A}(\eta)$.
 - Show that $R(\eta) \cong \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in R \right\}$.
 - If R is a division ring, show that $R(\eta)$ has exactly three ideals, 0 , $R\eta$ and R . [Hint: $a + b\eta$ is a unit if and only if $a \neq 0$.]
36. As in the discussion preceding Example 5 §3.2, define $R(\gamma)$ to be the set of all formal sums $a + b\gamma$, $a, b \in R$, where $\gamma^2 = 1$, $a\gamma = \gamma a$ for all $a \in R$, and $a + b\gamma = c + d\gamma$ if and only if $a = c$ and $b = d$.
- If A is an ideal of R , show that $A(\gamma)$ is an ideal of $R(\gamma)$ and $\frac{R(\gamma)}{A(\gamma)} \cong \frac{R}{A}(\gamma)$.
 - Show that $R(\gamma) \cong \left\{ \begin{bmatrix} a & b \\ b & a \end{bmatrix} \mid a, b \in R \right\}$.
 - If R is a division ring, show that $R(\gamma)$ has exactly four ideals, 0 , $R(1 + \gamma)$, $R(1 - \gamma)$ and $R(\gamma)$.
37. Show that $\mathbb{Z}_m \times \mathbb{Z}_n$ has a subring isomorphic to \mathbb{Z}_t , where $t = \text{lcm}(m, n)$.

38. If R^1 is as in Theorem 6, show that $R^1 \cong \mathbb{Z} \times R$. [Hint: Corollary to Theorem 7.]
39. Describe the maximal ideals in $R_1 \times R_2 \times \cdots \times R_n$, where $R_i \neq 0$ for each i . [Hint: Example 10.]
40. Let R be a ring in which $2 \in R^*$ and $\mu \in Z(R)$ exists such that $\mu^2 = -1$. Show that $R(i) \cong R \times R$. [Hint: Let $e = \frac{1}{2}(1 + ui)$ in the Corollary to Theorem 7.]
41. Let R be a ring in which $2 \in R^*$, and $u \in Z(R)$ exists such that $u^2 = \frac{1}{2}$. Show that $R(\sqrt{2}) \cong R \times R$. [Hint: Let $e = \frac{1}{2}(1 + u\sqrt{2})$ in the Corollary to Theorem 7.]
42. Let $\psi : R \rightarrow \frac{R}{A} \times \frac{R}{B}$ be the map given by $\psi(r) = (r + A, r + B)$ —see the proof of Theorem 8. If ψ is onto, show that necessarily $R = A + B$. [Hint: Choose r in R such that $\psi(r) = (1 + A, 0 + B)$.]
43. If X is a set and R is a ring, let $S = F(X, R)$ denote the ring of all mappings $X \rightarrow R$ using pointwise operations (see Example 4 §3.1).
- If R is a field and $x \in X$, show that $\{f \in S \mid f(x) = 0\}$ is a maximal ideal of S for each $x \in X$.
 - If M is a maximal ideal of R , show that $\{f \in S \mid f(x) \in M\}$ is a maximal ideal of S .
44. Let A_1, A_2, \dots, A_n be ideals of R and write $A = \bigcap_{i=1}^n A_i$.
- Show that R/A is isomorphic to a subring of $R/A_1 \times \cdots \times R/A_n$.
 - If $A_i + A_j = R$ for all $i \neq j$, show that $R/A \cong R/A_1 \times \cdots \times R/A_n$. [Hint: Show that $R = A_k + \left[\bigcap_{i \neq k} A_i \right]$ for each k by showing that this ideal contains 1. Let $1 = a_k + b_k$, $a_k \in A_k$, $b_k \in \bigcap_{i \neq k} A_i$. Given $(r_1 + A_1, \dots, r_n + A_n)$ in $\frac{R}{A_1} \times \cdots \times \frac{R}{A_n}$, consider $r = r_1 b_1 + \cdots + r_n b_n$.]

3.5 ORDERED INTEGRAL DOMAINS⁵⁷

The ring \mathbb{Z} of integers is an integral domain that has the additional property of being ordered: For m and n in \mathbb{Z} exactly one of $m < n$, $m = n$, or $n < m$ is true. There are other ordered integral domains (for example \mathbb{Q} or \mathbb{R}), but the integers have the further property that they are well ordered: Every set of positive integers has a smallest member. This assertion is the well-ordering axiom for \mathbb{Z} , which is equivalent to the principle of induction. The well-ordering axiom fails to hold for \mathbb{Q} or \mathbb{R} , and we devote this brief section to proving that it characterizes \mathbb{Z} among the ordered integral domains.

An integral domain R is said to be **ordered** if there is a subset $R^+ \subseteq R$, called the set of **positive elements** of R , satisfying the following conditions.

P1 *If a and b are in R^+ , then $a + b$ and ab are in R^+ .*

P2 *For all $a \in R$, exactly one of $a \in R^+$, $a = 0$, or $-a \in R^+$ holds.*

Write $a < b$ or $b > a$ to mean $b - a \in R^+$. Hence, \mathbb{Z} , \mathbb{Q} , and \mathbb{R} are ordered integral domains with the usual sets \mathbb{Z}^+ , \mathbb{Q}^+ , and \mathbb{R}^+ of positive elements. Note that we do

⁵⁷The material covered in this section is not needed elsewhere in the book.

not regard 0 as positive in \mathbb{Z} , \mathbb{Q} , or \mathbb{R} , and we retain this convention in any ordered integral domain R ($0 \notin R^+$ by P2).

Lemma 1. *Let $R \neq 0$ be an ordered integral domain.*

- (1) $R^+ = \{r \in R \mid r > 0\}$.
- (2) *If $a \in R$, exactly one of $a < 0$, $a = 0$, or $a > 0$ holds.*
- (3) *If $a < b$ and $b < c$ in R , then $a < c$.*
- (4) *If $a < b$ and $c > 0$ in R , then $ac < bc$.*
- (5) $a^2 > 0$ for all $a \neq 0$ in R . In particular, $1 > 0$.

Proof. (1) follows from the definition of $<$, and (2) restates P2. If $a < b$ and $b < c$, then $b - a$ and $c - b$ are in R^+ , so $c - a = (c - b) + (b - a)$ is also in R^+ by P1, proving (3). Similarly, (4) follows from P1 because $(b - a) \in R^+$ and $c \in R^+$ implies that $bc - ac = (b - a)c \in R^+$. As to (5), if $a \neq 0$, then $a > 0$ implies that $a^2 > 0$ by (4), whereas $a < 0$ implies that $-a > 0$, so again $a^2 = (-a)^2 > 0$. Finally, $1 \neq 0$ because $R \neq 0$, so $1 = 1^2 > 0$. \blacksquare

Lemma 1 shows that the complex numbers \mathbb{C} *cannot* be ordered. For if $\mathbb{C}^+ \subseteq \mathbb{C}$ satisfies P1 and P2, then $-1 = i^2 \in \mathbb{C}^+$ and $1 = 1^2 \in \mathbb{C}^+$ by (5), contradicting P2.

The well-ordering axiom (Section 1.1) is a potent property of the ring \mathbb{Z} of integers, as we have seen. The next theorem shows that it distinguishes \mathbb{Z} among the ordered domains. As for \mathbb{Z} , we say that an integral domain is **well ordered** if it is ordered and every nonempty set X of positive elements has a least member c (that is, $c \in X$ and $c < x$ for all x in X , $x \neq c$).

Theorem 1. *Let $R \neq 0$ be a well-ordered integral domain. Then an isomorphism $\sigma : \mathbb{Z} \rightarrow R$ exists such that, if $k < m$ in \mathbb{Z} , then $\sigma(k) < \sigma(m)$ in R .*

Proof. We begin with two preliminary results.

Claim 1. 1 is the least element of R^+ .

Proof. Let c be the least element of R^+ . Then one of $1 < c$, $c = 1$, or $c > 1$ must hold; $1 < c$ is ruled out because $1 \in R^+$. If $c < 1$, then $0 < c < 1$, so $0 < c^2 < c$ (by Lemma 1). Because $c^2 \in R^+$, this contradicts the minimality of c in R^+ . Hence, the only possibility is $c = 1$. This proves Claim 1.

Claim 2. $R^+ = \{k1 \mid k \in \mathbb{Z}^+\}$.

Proof. We first show that $k1 \in R^+$ for all $k \in \mathbb{Z}^+$ by induction on k . It is true if $k = 1$ because $1 \in R^+$. If $k1 \in R^+$ for some $k \in \mathbb{Z}^+$, then $(k+1)1 = k \cdot 1 + 1 \in R^+$ by P1, which proves that $\{k1 \mid k \in \mathbb{Z}^+\} \subseteq R^+$. If this is not equality, let d be the least member of $\{r \in R^+ \mid r \neq k1 \text{ for all } k \in \mathbb{Z}^+\}$. Because $d \in R^+$, either $d = 1$ or $1 < d$ by Claim 1. But $1 < d$ means $d - 1 \in R^+$ and $d - 1 < d$ (because $d - (d - 1) = 1 \in R^+$). Thus, the choice of d implies that $d - 1 = k1$ for some $k \in \mathbb{Z}^+$, and so $d = k1 + 1 = (k+1)1$, a contradiction. This proves Claim 2.

We can now prove Theorem 1. Define $\sigma : \mathbb{Z} \rightarrow R$ by $\sigma(k) = k1$. Then we have $\sigma(k+m) = \sigma(k) + \sigma(m)$ and $\sigma(km) = \sigma(k) \cdot \sigma(m)$ for all $k, m \in \mathbb{Z}$ (see Theorem 2 §3.1), and $k < m$ implies $\sigma(k) < \sigma(m)$ because $\sigma(m) - \sigma(k) = (m - k)1 \in R^+$ by Claim 2.

To prove that σ is one-to-one, let $\sigma(k) = \sigma(m)$. Then $(k - m)1 = 0 \notin R^+$, so $k \leq m$ by Claim 2. But $(m - k)1 = -(k - m)1 = 0$ too, so $k \geq m$. Hence, $k = m$ and σ is one-to-one.

Finally, σ is onto. If $r \in R$, there are three cases: $r = 0$, $r > 0$, and $r < 0$. If $r = 0$, then $r = \sigma(0)$; if $r > 0$, then $r = \sigma(k)$ for some $k \in \mathbb{Z}^+$ by Claim 2; if $r < 0$, then $-r > 0$, so $r = \sigma(-k)$ for $k \in \mathbb{Z}^+$. Hence, σ is onto as required. ■

Exercises 3.5

1. Let R be an ordered integral domain and let a, b , and c denote elements of R . Show that
 - (a) If $a < b$, then $a + c < b + c$ for all $c \in R$.
 - (b) If $a < b$ and $c < 0$, then $ac > bc$.
 - (c) If $a < b$, then $-a > -b$.
 - (d) If $a < b$ and $c < d$, then $a + c < b + d$.
 - (e) If $0 < a < b$ and $0 < c < d$, then $ac < bd$.
 - (f) If $ab < ac$ and $a > 0$, then $b < c$.
 2. Write $a \leq b$ in an ordered integral domain R to mean $a < b$ or $a = b$. Show that
 - (a) $a \leq a$ for all $a \in R$.
 - (b) If $a \leq b$ and $b \leq a$, then $a = b$.
 - (c) If $a \leq b$ and $b \leq c$, then $a \leq c$.

Because of this, \leq is called a **partial order** on R .
 3. If R is an ordered integral domain, define the **absolute value** $|a|$ of $a \in R$ by
- $$|a| = \begin{cases} a, & \text{if } 0 \leq a, \\ -a, & \text{if } a < 0. \end{cases}$$
- Prove the following for all a and b in R .
- (a) $|a| \geq 0$
 - (b) $-|a| \leq a \leq |a|$
 - (c) $|ab| = |a||b|$
 - (d) $|a + b| \leq |a| + |b|$
4. If R is an ordered integral domain and $a \in R$, show that $b \in R$ exists such that $a < b$. Conclude that R has no largest member.
 5. In each case, show that the integral domain R cannot be ordered.
 - (a) $\mathbb{Z}(i)$ —the gaussian integers
 - (b) \mathbb{Z}_p , p a prime
 6. Suppose that $u > 0$ and $u^2 = 2$ in an ordered integral domain R . Prove that $2u < 3$, where $2 = 1 + 1$ and $3 = 2 + 1$.
 7. Let R be an ordered integral domain and let Q denote the field of quotients of R . Show that Q is ordered if $Q^+ = \{r/u \mid ru \in R^+\}$.

Chapter 4

Polynomials

One cannot escape the feeling that these mathematical formulae have an independent existence and an intelligence of their own, that they are wiser than we are, wiser even than their discoverers, that we get more out of them than was originally put into them.

—Heinrich Hertz

The study of polynomials is the oldest branch of algebra. The Hindus knew how to solve quadratics in 600 BC, and the Babylonians by then had developed considerable skill at algebraic manipulation and were using special cases of the quadratic formula. However, symbolic algebra in the form we know it today developed in Arabia between 600 and 1000 AD. They were solving cubic equations and, in the work of al-Khowarizmi (c.825), were starting to identify geometric magnitudes with numbers. These efforts led them to the familiar formulas for areas, volumes, and the like. By Descartes' time (1596–1650), analytic geometry was well understood, so that the computational power of algebra and the intuitive power of geometry could each enhance the other.

Subsequently, the *theory of equations* attracted the best mathematicians. Euler and Lagrange considered the problem of finding a general formula, analogous to the quadratic formula, for the roots of any quintic polynomial (degree 5). Their work led to the epoch making discovery of Abel who, in 1823, showed that no such formula exists. Later Galois showed it is impossible for any polynomial of degree 5 or more, and brought groups into the picture.

The general study of curves and surfaces as graphs of polynomials is known as algebraic geometry. A central problem here is to discover which properties of a curve or a surface remain invariant under certain transformations given by polynomials in the coordinates. This *invariant theory* dates from Cayley's time (1821–1895) and continues to be an active research area today.

4.1 POLYNOMIALS

The reader is doubtless acquainted with polynomials, having had to graph equations such as $y = x^2 - 2x - 2$, obtain factorizations such as $6x^2 - 11x + 3 = (2x - 3)(3x - 1)$, and find solutions (called roots) of equations such as $x^2 - 2x - 2 = 0$. Moreover, polynomials are associated with geometry. For example, the graph of $y = 3x - 2$ is a line and the graph of $y = 5x^2 - x + 7$ is a parabola. In addition, polynomials are treated as formulas for functions. For example, a function $f : \mathbb{R} \rightarrow \mathbb{R}$ could be defined by $f(x) = x^3 - x - 1$. In fact many readers will already know how to differentiate and integrate such polynomial functions.

If R is any ring, a symbol x is called an **indeterminate** over R if

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = 0, \quad a_i \in R, \quad \text{implies} \quad a_i = 0 \text{ for each } i.$$

The study of polynomials requires the existence of such an element for any ring R . Clearly, if $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ is to make sense, x and the a_i must belong to some ring, and we begin by constructing such a ring.

Lemma 1. Given a ring R , there exists a ring S with the following properties:

- (1) $R \subseteq S$ is a subring.
- (2) There exists $x \in S$ such that x is an indeterminate over R .
- (3) If $a \in R$ then $ax = xa$.

Proof. We sketch the proof; the details are in Section 4.6. A function $\alpha : \mathbb{N} \rightarrow R$ is called a **sequence** from R . If we write $\alpha(k) = a_k$ for each $k \geq 0$, we denote this sequence by $\alpha = [a_k] = [a_0, a_1, a_2, \dots]$. Given another sequence $\beta = [b_k]$, we have $[a_k] = [b_k]$ if and only if $\alpha = \beta$, that is if and only if $a_k = b_k$ for all $k \geq 0$.

If S is the set of all sequences from R , then S becomes a ring if we define

$$[a_k] + [b_k] = [a_k + b_k],$$

$$[a_k][b_k] = [p_k], \quad \text{where } p_k = a_0b_k + a_1b_{k-1} + \cdots + a_{k-1}b_1 + a_kb_0 \text{ for } k \geq 0.$$

Moreover, R is a subring of S if we identify $a = [a, 0, 0, 0, \dots]$ for $a \in R$, proving (1). This being done, define $x = [0, 1, 0, 0, \dots] \in S$. Then, with some calculation, we obtain $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = [a_0, a_1, a_2, \dots, a_n, 0, 0, \dots]$ for all $a_i \in R$, and (2) follows. Finally $ax = [0, a, 0, \dots] = xa$ for all $a \in R$, proving (3). ■

Let x be an indeterminate over a ring R , and let S be as in Lemma 1. Then

$$R[x] = \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \mid n \geq 0, a_i \in R \text{ for each } i\}.$$

is a subring of S , called the **ring of polynomials** over R . Here an expression $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ in $R[x]$ is called a **polynomial** over R . The elements a_i in R are called the **coefficients** of f , and they are uniquely determined by f (x is an indeterminate). The polynomial f will often be written simply as

$$f = a_0 + a_1x + \cdots$$

where it is understood that the sum is finite, that is all coefficients are zero from some point on. Two polynomials

$$f = a_0 + a_1x + a_2x^2 + \cdots \quad \text{and} \quad g = b_0 + b_1x + b_2x^2 + \cdots$$

in the ring $R[x]$ are **equal** (and we write $f = g$) if and only if $f - g = 0$, that is $(a_0 - b_0) + (a_1 - b_1)x + \dots = 0$. Because x is an indeterminate, this means

$$f = g \quad \text{if and only if} \quad a_k = b_k, \quad \text{for all } k = 0, 1, 2, \dots$$

Hence, for example, we cannot write $2x^2 - 3x + 1 = 0$ in $\mathbb{R}[x]$ because it would mean $2 = 0$, $-3 = 0$, and $1 = 0$. Instead we refer to finding a **root** in \mathbb{R} of the function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = 2x^2 - 3x + 1$, that is an element $a \in \mathbb{R}$ such that $0 = f(a) = 2a^2 - 3a + 1$. In this case $a = 1$ and $a = \frac{1}{2}$ are the only possibilities.

Note that when writing a polynomial we omit terms $0x^i$ where the coefficient is zero. For example we write $2x - 3x^3$ rather than $0 + 2x + 0x^2 - 3x^3$. The coefficient a_0 (which may be 0) is called the **constant coefficient** of $f = a_0 + a_1x + \dots$. If all the other coefficients are zero, $f = a_0$ is called a **constant polynomial**, and the ring R is the subring of *all* constant polynomials in $R[x]$.

The **zero** and **unity** of the ring $R[x]$ are the constant polynomials 0 and 1, respectively. The **negative** of a polynomial $f = a_0 + a_1x + a_2x^2 + \dots$ is the polynomial $-f = -a_0 - a_1x - a_2x^2 - \dots$ where we negate every coefficient of f .

We single out the following facts for reference:

Lemma 2. Let $f = a_0 + a_1x + a_2x^2 + \dots$ and $g = b_0 + b_1x + b_2x^2 + \dots$ be polynomials in $R[x]$ where R is any ring. Then

- (1) $f = g$ if and only if $a_i = b_i$ for each i .
- (2) $f + g = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \dots$.
- (3) $fg = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \dots$.
- (4) The coefficient of x^k in fg is $a_0b_k + a_1b_{k-1} + \dots + a_{k-1}b_1 + a_kb_0 = \sum_{i+j=k} a_i b_j$. ■

Example 1. If $f = a_0 + a_1x + a_2x^2$ and $g = b_0 + b_1x$, then

$$fg = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_1b_1 + a_2b_0)x^2 + a_2b_1x^3.$$

Example 2. In $\mathbb{Z}[x]$, $(1 - 2x + x^3)(2 - x + x^2) = 2 - 5x + 3x^2 - x^4 + x^5$.

Example 3. In $\mathbb{Z}_3[x]$, $(x + 1)^3 = x^3 + 3x^2 + 3x + 1 = x^3 + 1$ because $3 = 0$ in $\mathbb{Z}_3[x]$.

The following theorem summarizes the above discussion.

Theorem 1. Let R be a ring and let x be an indeterminate over R . Then:

- (1) $R[x]$ is a ring.
- (2) R is the subring of all constant polynomials in $R[x]$.
- (3) If $Z = Z(R)$ denotes the center of R , then the center of $R[x]$ is $Z[x]$.
- (4) In particular, x is in the center of $R[x]$.
- (5) If R is commutative, then $R[x]$ is commutative.

Proof. We already have (1) and (2), and (3) implies (4) and (5). So we prove (3).

(3). To see that $Z[x] \subseteq Z(R[x])$, let $f = z_1 + z_1x + z_2x^2 + \dots \in Z[x]$, that is each $z_i \in Z$. Then $z_i b = bz_i$ for every $b \in R$, so Lemma 2 implies that $fg = gf$ for each polynomial $g \in R[x]$. It follows that $Z[x] \subseteq Z(R[x])$. Conversely, suppose that $f = a_0 + a_1x + a_2x^2 + \dots$ is central in $R[x]$. Then $fa = af$ for every $a \in R$, so Lemma 2 shows that $a_i a = aa_i$. This means each $a_i \in Z$, so $Z(R[x]) \subseteq Z[x]$. ■

Lemma 3. Let $\langle x \rangle = \{xf \mid f \in R[x]\}$ denote the set of all multiples of x in $R[x]$. Then $\langle x \rangle$ is an ideal of $R[x]$, and $R[x]/\langle x \rangle \cong R$.

Proof. Define $\theta : R[x] \rightarrow R$, by $\theta(a_0 + a_1x + \dots) = a_0$. This is well defined by Lemma 2, and satisfies $\theta(1) = 1$. If $f = a_0 + a_1x + \dots$ and $g = b_0 + b_1x + \dots$ in $R[x]$, the constant coefficients of $f + g$ and fg are $a_0 + b_0$ and a_0b_0 , respectively. This means that θ is an onto ring homomorphism. Since $\ker \theta = \langle x \rangle$, it follows by the isomorphism theorem that $\langle x \rangle$ is an ideal of $R[x]$ and $R[x]/\langle x \rangle \cong R$. ■

If f is any nonzero polynomial in $R[x]$, the highest exponent of x appearing in f (with nonzero coefficient) is called the **degree** of f , written $\deg f$; the coefficient itself is called the **leading coefficient** of f . If the leading coefficient is 1, f is called **monic**. The degree of the zero polynomial is not defined. Polynomials of degree 1, 2, 3, 4, and 5 are called, respectively, **linear**, **quadratic**, **cubic**, **quartic**, and **quintic** polynomials.

Example 4. The polynomial $x - x^2 + 2x^3$ has degree 3, $x + 2$ is monic of degree 1, and -5 has degree 0. The polynomials in $R[x]$ of degree 0 are just the nonzero constant polynomials, that is the nonzero elements of R .

Suppose $f \neq 0$ and $g \neq 0$ are polynomials in $R[x]$, say

$$f = a_0 + a_1x + a_2x^2 + \dots + a_mx^m \quad \text{and} \quad g = b_0 + b_1x + b_2x^2 + \dots + b_nx^n,$$

where $a_m \neq 0$ and $b_n \neq 0$. Thus, $\deg f = m$ and $\deg g = n$, and a_m and b_n are the leading coefficients. Clearly,

$$fg = a_0b_0 + (a_0b_1 + a_1b_0)x + \dots + (a_mb_n)x^{m+n}.$$

It is possible that $fg = 0$, but if $fg \neq 0$ it follows that $\deg(fg) \leq \deg f + \deg g$. However, if R is a domain, then $a_mb_n \neq 0$ and it follows that

$$fg \neq 0 \quad \text{and} \quad \deg fg = m + n = \deg f + \deg g.$$

This proves (1) and (2) of the following theorem.

Theorem 2. Let R be a domain. Then:

- (1) $R[x]$ is a domain.
- (2) If $f \neq 0$ and $g \neq 0$ in $R[x]$, then $\deg(fg) = \deg f + \deg g$.
- (3) The units in $R[x]$ are the units in R .

Proof. (1) and (2) are proved in the above discussion.

(3). If f is a unit in $R[x]$, denote its inverse by g . Then $fg = 1 = gf$, so (2) gives $\deg f + \deg g = \deg 1 = 0$. But $\deg f$ and $\deg g$ are nonnegative integers, so this implies that $\deg f = 0 = \deg g$. Hence f and g are (nonzero) elements of R , so f is a unit in R . Conversely, each unit u in R is a unit in $R[x]$ with inverse the constant polynomial u^{-1} . ■

The next example shows that it is vital that R is a domain in Theorem 2.

Example 5. Consider $f = 1 + 2x$ in $\mathbb{Z}_4[x]$. Then the fact that $4 = 0$ in \mathbb{Z}_4 gives

$$f^2 = (1 + 2x)(1 + 2x) = 1 + (2 + 2)x + 2^2x^2 = 1.$$

Hence f is a (self-inverse) unit in $\mathbb{Z}_4[x]$ that is not in \mathbb{Z}_4 , so part (3) of Theorem 2 fails in $\mathbb{Z}_4[x]$. Moreover, (2) also fails because $\deg(f^2) = \deg 1 = 0$ whereas $\deg f + \deg f = 1 + 1 = 2$. Finally, (1) fails because $(2x)^2 = 0$ in $\mathbb{Z}_4[x]$. \square

On the other hand, if F is a field it can be shown that $F[x] \times F[x] \cong (F \times F)[x]$, so $R = F \times F$ is a nondomain for which the units in $R[x]$ are in R .

The proof of (1) and (2) in Theorem 2 extends to another important case where the degree function behaves well (Theorem 3 below). The proof is Exercise 7.

Theorem 3. Let R be any ring and let $f \neq 0$ and $g \neq 0$ be polynomials in $R[x]$. If the leading coefficient of either f or g is a unit in R , then

- (1) $fg \neq 0$ in $R[x]$.
- (2) $\deg(fg) = \deg f + \deg g$.

The Division Algorithm

Our discussion of the factorization of integers in Section 1.2, and of the ring \mathbb{Z}_n of integers modulo n in Section 1.3, both depend in a fundamental way on the division algorithm (Theorem 1 §1.2): Given m and $n > 0$ in \mathbb{Z} , uniquely defined integers q and r exist such that $m = qn + r$ and $0 \leq r < n$. The standard process of “long division” is an algorithm for computing q and r , and an analogous procedure works for polynomials, as shown in Example 6.

Example 6. Given $f = x^2 + 1$ and $g = x^4 + 3x^3 + x + 1$ in $\mathbb{Z}[x]$, find polynomials q and r such that

$$g = qf + r, \text{ and either } r = 0 \text{ or } \deg r < \deg f = 2.$$

Solution. The following tableau describes the process.

$$\begin{array}{r} x^2 + 3x - 1 \\ \hline x^2 + 1 | x^4 + 3x^3 + x + 1 \\ x^4 + 0x^3 + 0x^2 + 0x + 1 \\ \hline 0x^4 + 3x^3 + x^2 + x + 1 \\ 0x^4 + 3x^3 + 0x^2 + 0x + 1 \\ \hline 0x^3 + x^2 + 2x + 1 \\ 0x^3 + x^2 + 0x + 1 \\ \hline 0x^2 + 2x + 1 \\ 0x^2 + 2x + 0 \\ \hline 0x + 1 \\ 0x + 2 \end{array}$$

Hence, $q = x^2 + 3x - 1$ and $r = -2x + 2$ in this case. The reader should verify that $g = qf + r$ really is true.

The quotient q appears at the top and is created one term at a time from left to right. At each stage we choose the new term in q so that, when multiplied by the divisor $x^2 + 1$, the result has the same leading coefficient as the last polynomial in the tableau at that stage. The process stops when this operation cannot be achieved, that is, when the last polynomial in the tableau is either 0 or has degree less than the degree of the divisor (in this case, less than 2). This last polynomial is the remainder $r = -2x + 2$ above. \square

This division process requires that the leading coefficient of the divisor is a unit; in fact, in most cases of interest the divisor is actually monic (as in Example 6). Apart from this requirement, the algorithm works in complete generality and the proof (by induction) is an adaptation of the algorithm itself.

Theorem 4. Division Algorithm. Let R be any ring and let f and g be polynomials in $R[x]$. Assume that $f \neq 0$ and that the leading coefficient of f is a unit in R . Then uniquely determined polynomials q and r exist in $R[x]$ such that

- (1) $g = qf + r$.
- (2) Either $r = 0$ or $\deg r < \deg f$.

Proof. We first prove that such q and r exist. Write $m = \deg g$ and $n = \deg f$. If $g = 0$ or $m < n$, then $g = 0f + g$ does it. So assume that $m \geq n$ and proceed by induction on m . Write $f = ux^n + ax^{n-1} + \dots$ and $g = bx^m + cx^{m-1} + \dots$, where u is a unit in R by hypothesis. Consider the new polynomial

$$\begin{aligned} g_1 &= g - bu^{-1}x^{m-n}f \\ &= (bx^m + cx^{m-1} + \dots) - bu^{-1}x^{m-n}(ux^n + ax^{n-1} + \dots) \\ &= 0x^m + (c - bu^{-1}a)x^{m-1} + \dots, \end{aligned}$$

where we used the fact that x is central in $R[x]$. Hence either $g_1 = 0$ or $\deg g_1 < m$ so, by induction, polynomials q_1 and r exist such that $g_1 = q_1 f + r$ and either $r = 0$ or $\deg r < \deg f = n$. But then

$$g = g_1 + bu^{-1}x^{m-n}f = (q_1 f + r) + bu^{-1}x^{m-n}f = (q_1 + bu^{-1}x^{m-n})f + r.$$

This completes the induction, so q and r exist satisfying (1) and (2).

To prove uniqueness, suppose that also $g = q_1 f + r_1$, where either $r_1 = 0$ or $\deg r_1 < \deg f$. Then $r - r_1 = (q_1 - q)f$. If $q_1 - q \neq 0$ then, since the leading coefficient of f is a unit, Theorem 3 implies that $(q_1 - q)f \neq 0$ and that

$$\deg(r - r_1) = \deg[(q_1 - q)f] = \deg(q_1 - q) + \deg f.$$

But this implies that $\deg(r - r_1) \geq \deg f$, a contradiction. So, $q_1 - q = 0$, whence $r - r_1 = (q_1 - q)f = 0$. This proves the uniqueness. \blacksquare

We use the division algorithm repeatedly. However, even though we proved it for an arbitrary ring R , it is most effective when R is commutative. The reason is as follows. Given $a \in R$ and a polynomial $f = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$ in $R[x]$, we want to substitute a for x to get $f(a) = a_0 + a_1a + a_2a^2 + \dots + a_na^n$. But we must be careful when doing this.

To illustrate, suppose that a and b are given in a (possibly noncommutative) ring R , and consider the polynomial

$$f = (x - a)(x + b).$$

Then we appear to have

$$f(a) = (a - a)(a + b) = 0(a + b) = 0.$$

However, $f = x^2 + (b - a)x - ab$ when multiplied out, so

$$f(a) = a^2 + (b - a)a - ab = ba - ab.$$

This is clearly nonsense if $ab \neq ba$. The reason is that, in our construction of the ring $R[x]$, we insisted that $rx = xr$ holds for all $r \in R$. So substituting a for x in f is bound to create problems unless a commutes with all the coefficients of f . Hence, if substituting $x = a$ is to make sense for all polynomials in $R[x]$, the element a must be in the center of R . However, in this case substitution works well.

Theorem 5. Evaluation Theorem. Let R be a ring and let a be an element in the center $Z(R)$ of R . Define a mapping $\varphi_a : R[x] \rightarrow R$ by

$$\varphi_a(a_0 + a_1x + a_2x^2 + \cdots + a_nx^n) = a_0 + a_1a + a_2a^2 + \cdots + a_na^n.$$

Then φ_a is an onto ring homomorphism.

Proof. For a constant polynomial c , have $\varphi_a(c + 0x + 0x^2 + \cdots) = c$, so φ_a is onto and $\varphi_a(1) = 1$. To show that φ_a preserves addition and multiplication, let

$$f = a_0 + a_1x + a_2x^2 + \cdots \quad \text{and} \quad g = b_0 + b_1x + b_2x^2 + \cdots$$

be polynomials in $R[x]$. Then $f + g = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots$ so

$$\begin{aligned}\varphi_a(f + g) &= (a_0 + b_0) + (a_1 + b_1)a + (a_2 + b_2)a^2 + \cdots \\ &= (a_0 + a_1a + a_2a^2 + \cdots) + (b_0 + b_1a + b_2a^2 + \cdots) \\ &= \varphi_a(f) + \varphi_a(g)\end{aligned}$$

Hence, φ_a preserves addition. Turning to multiplication, recall that

$$fg = c_0 + c_1x + c_2x^2 + \cdots$$

where the coefficient c_k of x^k is given by $c_k = a_0b_k + a_1b_{k-1} + \cdots + a_kb_0$ for each $k \geq 0$ (see Lemma 2). Because a is central in R , we have

$$\begin{aligned}\varphi_a(f)\varphi_a(g) &= (a_0 + a_1a + a_2a^2 + \cdots)(b_0 + b_1a + b_2a^2 + \cdots) \\ &= a_0b_0 + (a_0b_1a + a_1ab_0) + (a_0b_2a^2 + a_1ab_1a + a_2a^2b_0) + \cdots \\ &= a_0b_0 + (a_0b_1 + a_1b_0)a + (a_0b_2 + a_1b_1 + a_2b_0)a^2 + \cdots \\ &= c_0 + c_1a + c_2a^2 + \cdots \\ &= \varphi_a(fg).\end{aligned}$$

Thus φ_a preserves multiplication, and so is a ring homomorphism. ■

Let R be a ring, let $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ be a polynomial in $R[x]$, and let $a \in Z(R)$ be central in R . We write $f(a) = a_0 + a_1a + \cdots + a_na^n$ for the element of R obtained by substituting a for x . Then $f(a)$ is called the **evaluation** of f at a . For example, if $f = 5 + 4x - 2x^2 + x^3 \in \mathbb{Z}[x]$ then $f(3) = 5 + 4 \cdot 3 - 2 \cdot 9 + 27 = 26$.

Example 7. Consider $f = x^3 - x$ in $\mathbb{Z}_6[x]$. Show that $f(a) = 0$ for all $a \in \mathbb{Z}_6$.

Solution. One verifies that $a^3 = a$ for all $a \in \mathbb{Z}_6$. Since \mathbb{Z}_6 is commutative, we have $f(a) = a^3 - a = 0$ for all $a \in \mathbb{Z}_6$. □

Example 8. If $f = c$ is a constant polynomial, then $f(a) = c$ for all $a \in Z(R)$.

If $a \in Z(R)$, the homomorphism $\varphi_a : R[x] \rightarrow R$ in Theorem 5 is called the **evaluation map** because

$$\varphi_a(f) = f(a), \quad \text{for all } f \in R[x].$$

The gist of Theorem 5 is that evaluation at a satisfies

$$(fg)(a) = f(a)g(a) \quad \text{and} \quad (f + g)(a) = f(a) + g(a), \quad \text{for all } f, g \in R[x].$$

Perhaps the most useful consequence of evaluation is

Example 9. If $f = gh$ in $R[x]$, and if $g(a) = 0$ where $a \in Z(R)$, then $f(a) = 0$ too. In particular, if $f = (x - a)h$ where $h \in R[x]$, then $f(a) = 0$.

Solution. Since a is central in R , Theorem 5 gives $f(a) = g(a)$. If $h(a) = 0$, then $f(a) = (a - a)h(a) = 0$. If $f = (x - a)h$ then $f(a) = (a - a)h(a) = 0$. \square

In Lemma 3, we showed directly that the mapping $\theta : R[x] \rightarrow R$ is a ring homomorphism, where $\theta(f)$ is the constant term of f . However, this follows easily from Theorem 5 because $\theta = \varphi_0$ is evaluation at the central element 0. Here is another example.

Example 10. Define $\theta : R[x] \rightarrow R$ where $\theta(f)$ is the sum of the coefficients of f . Show that θ is a ring homomorphism.

Solution. If $f = a_0 + a_1x + a_2x^2 + \dots$, then $\theta(f) = a_0 + a_1 + a_2 + \dots = f(1)$. Thus $\theta = \varphi_1$ is evaluation at 1, and so is a homomorphism because 1 is central. \square

Commutative Rings

Evaluation is most useful when R is commutative, so we make that assumption for the rest of this chapter. In this case, there is a special notation that is commonly used. If $a \in R$ and R is commutative, the set $Ra = \{ra \mid r \in R\}$ of all multiples of a is an ideal of R , called the **principal ideal generated by a** , and denoted

$$\langle a \rangle = Ra = \{ra \mid r \in R\}.$$

We use this notation in the (commutative) ring $R[x]$.

Theorem 6. Let R be commutative ring, let $a \in R$, and let $f \in R[x]$. Then

- (1) **Factor Theorem.** $f(a) = 0$ if and only if $f = (x - a)q$ for some $q \in R[x]$; that is, if and only if $f \in \langle x - a \rangle$.
- (2) **Remainder Theorem.** If f is divided by $x - a$, the remainder is $f(a)$.

Proof. Write $f = (x - a)q + r$ by the division algorithm where q and r are in $R[x]$ and either $r = 0$ or $\deg r < \deg(x - a) = 1$. In both cases $r \in R$ is a constant polynomial. Now evaluation at a gives $f(a) = (a - a)q(a) + r = r$. This proves (2), and also shows that $f = (x - a)q + f(a)$. But then $f(a) = 0$ if and only if $f = (x - a)q$ for some $q \in R[x]$, proving (1). \blacksquare

Corollary. If R is commutative, let $\varphi_a : R[x] \rightarrow R$ be evaluation at $a \in R$. Then $\ker \varphi_a = \langle x - a \rangle$ and $R[x]/\langle x - a \rangle \cong R$.

Proof. We have $\ker \varphi_a = \{f \in R[x] \mid f(a) = 0\} = \langle x - a \rangle$ by the factor theorem. Now apply the isomorphism theorem. \blacksquare

Example 11. If $f = 2x^3 + x + 1 \in \mathbb{Z}_6[x]$, verify the remainder theorem for $a = 2$.

Solution. The division algorithm gives $f = (x - 2)(2x^2 + 4x + 3) + 1$ (see the tableau), so the remainder is $1 = f(2)$, as required.

$$\begin{array}{r} 2x^2 + 4x + 3 \\ x - 2 \quad \boxed{2x^3} + x + 1 \\ \hline 2x^3 + 2x^2 \\ \hline 4x^2 + x + 1 \\ 4x^2 + 4x \\ \hline 3x + 1 \\ 3x \\ \hline 1 \end{array}$$

\square

Let $f \in R[x]$, where R is commutative. Then $a \in R$ is called a **root** of f if it satisfies the following conditions (equivalent by the factor theorem):

- (1) $f(a) = 0$.
- (2) $f = (x - a)q$ for some $q \in R[x]$.
- (3) $f \in \langle x - a \rangle$.

Thus, every element of R is a root of the zero polynomial, while a nonzero constant polynomial has no roots.

Suppose $f \neq 0$ has degree n . If a is a root of f then $f = (x - a)q_1$ in $R[x]$ by the factor theorem. Moreover, $\deg q_1 = n - 1$ by Theorem 3. If $q_1(a) = 0$, another application of the factor theorem gives $f = (x - a)^2 q_2$ where $\deg q_2 = n - 2$. If $q_2(a) = 0$, the process continues. Because the degrees of the quotients q_1, q_2, \dots decrease, the process must end with $q_m(a) \neq 0$ for some $m > 0$. This leads to the following terminology: If $f = (x - a)^m q$, where $q \in R[x]$ and $q(a) \neq 0$, the root a is said to have **multiplicity** $m \geq 1$.

Example 12. In $\mathbb{Z}_8[x]$, find the multiplicity of 2 as a root of the polynomial $f = x^4 + 5x^3 + 3x^2 + 4$.

Solution. We have $f = (x - 2)q_1$ by the division algorithm, where $q_1 = x^3 - x^2 + x + 2$. But $q_1(2) = 0$ too, so $q_1 = (x - 2)q_2$, where $q_2 = x^2 + x + 3$. As $q_2(2) \neq 0$, the multiplicity is 2 and $f = (x - 2)^2(x^2 + x + 3)$. \square

Examples 13 and 14 show how the number of roots of a polynomial depends on the ring.

Example 13. The polynomial $x^2 + 1$ has no roots in \mathbb{R} but two in \mathbb{C} , i and $-i$.

Example 14. Consider the polynomial $x^2 - 1$. It has roots 1 and -1 in any commutative ring, 1 and -1 are the *only* roots in any integral domain (verify), and it has four roots 1, 3, 5, and 7 in \mathbb{Z}_8 . If $S = \mathbb{Z}_2[x]$ write

$$R = \left\{ \begin{bmatrix} r & s \\ 0 & r \end{bmatrix} \mid r \in \mathbb{Z}_2, s \in S \right\} \quad \text{and} \quad \sigma = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix}.$$

Then one verifies that R is a commutative ring and $\sigma^2 = 1$ for any $s \in S$. Hence, $x^2 - 1$ has infinitely many roots in R (because S is infinite).

Examples 13 and 14 indicate that not much can be said in general about the number of roots of a polynomial over a commutative ring. However, if the ring is an integral domain we do have Theorem 8.

Theorem 8. Let R be an integral domain and let f be a nonzero polynomial in $R[x]$ of degree n . Then f has at most n roots in R .

Proof. Use induction on $n = \deg f$. If $n = 0$, then f is a nonzero constant and has no roots. If $n = 1$, say $f = a_0 + a_1x$ where $a_1 \neq 0$, let a and b be roots of f . Then $a_0 + a_1a = 0 = a_0 + a_1b$, so $a_1(b - a) = 0$. Hence, $b = a$ because R is a domain.

Suppose $n > 1$. If f has no root in R , we are done. If $f(a) = 0$, with $a \in R$, then $f = (x - a)q$ by the factor theorem, and $\deg q = n - 1$ by Theorem 2. Suppose that $b \neq a$ is another root of f . Then $0 = f(b) = (b - a)q(b)$, so $q(b) = 0$ because R is a domain. But q has at most $n - 1$ roots in R by induction, so f has at most $n - 1$ roots distinct from a . Hence f has at most n roots, as required. \blacksquare

We hasten to note that a polynomial of degree n over an integral domain R need not have *any* roots in R (for example $x^2 + 1$ where $R = \mathbb{Z}$). The force of Theorem 8 is to place a *maximum* on the number of roots. Also, it is important that R is commutative: $x^2 + 1$ has roots $\pm i$, $\pm j$, and $\pm k$ in the division ring \mathbb{H} of quaternions.

In factoring a polynomial such as $f = 6x^2 - 7x + 2 = (2x - 1)(3x - 2)$ in $\mathbb{Z}[x]$, it is important to be able to find the *rational* roots of f , that is the roots in \mathbb{Q} . Theorem 9 reduces this task to checking a finite number of potential roots.

Theorem 9. Rational Roots Theorem. Let $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ be a polynomial in $\mathbb{Z}[x]$, where $a_0 \neq 0$ and $a_n \neq 0$. Then every rational root of f has the form $\frac{c}{d}$ where $c|a_0$ and $d|a_n$.

Proof. If $\frac{c}{d} \in \mathbb{Q}$ is a root of f we may assume that it is in lowest terms; that is, $\gcd(c, d) = 1$. We have $0 = f\left(\frac{c}{d}\right) = a_0 + a_1\frac{c}{d} + \cdots + a_n\frac{c^n}{d^n}$. Multiplying by d^n gives

$$0 = a_0d^n + a_1cd^{n-1} + \cdots + a_{n-1}c^{n-1}d + a_nc^n.$$

Then $c|a_0d^n$ because c appears in each term but the first. But $\gcd(c, d^n) = 1$ (any prime dividing c and d^n would divide c and d), so Theorem 5 §1.2 gives $c|a_0$. A similar argument shows that $d|a_n$. ■

Corollary. The only rational roots of a monic polynomial in $\mathbb{Z}[x]$ are integers.

Example 15. Let m be a positive integer that is not the square of an integer. Show that \sqrt{m} is not in \mathbb{Q} .

Solution. If \sqrt{m} were in \mathbb{Q} , it would be a rational root of $x^2 - m$. If $q = \frac{c}{d}$ is such a root (in lowest terms), Theorem 9 shows that $c|m$ and $d|1$. Hence $d = \pm 1$, so $q = \pm c \in \mathbb{Z}$. Neither of these are roots of $x^2 - m$ by hypothesis. □

Example 16. Factor $f = 3x^3 - x^2 - x - 4$ as far as possible in $\mathbb{Q}[x]$.

Solution. If $\frac{c}{d}$ is a rational root of f , then $c|(-4)$ and $d|3$ by Theorem 9. Hence $c = \pm 1, \pm 2, \pm 4$ and $d = \pm 1, \pm 3$; so $\frac{c}{d} = \pm 1, \pm 2, \pm 4, \pm \frac{1}{3}, \pm \frac{2}{3}, \pm \frac{4}{3}$. Exhaustive checking gives $\frac{4}{3}$ as a root, so $x - \frac{4}{3}$ is a factor in $\mathbb{Q}[x]$. Hence $3x - 4$ is a factor in $\mathbb{Q}[x]$ and the division algorithm gives

$$f = (3x - 4)(x^2 + x + 1) \text{ in } \mathbb{Q}[x].$$

But $x^2 + x + 1$ has no rational roots by Theorem 9 (the possibilities are ± 1), so $x^2 + x + 1$ has no factorization in $\mathbb{Q}[x]$ (any factors would be linear and so produce rational roots). Hence f factors no further in $\mathbb{Q}[x]$. □

Note that we are not done factoring in Example 16 if we allow the factors to have coefficients in \mathbb{C} (because $x^2 + x + 1$ has roots $\frac{1}{2}[-1 \pm \sqrt{3}i]$ in \mathbb{C} by the quadratic formula). Note also that f actually factored in $\mathbb{Z}[x]$, even though it has no root in \mathbb{Z} . We examine these observations further in Section 4.2.

Exercises 4.1

Throughout these exercises R is a ring unless otherwise specified.

1. In each case compute $f + g$ and fg .
 - (a) $f = 3 + 2x + x^2 + 4x^3$, $g = 1 + x^2 + x^3$ in $\mathbb{Z}_5[x]$
 - (b) $f = 5 + 2x + x^2 + x^3$, $g = 2 + x + x^2$ in $\mathbb{Z}_7[x]$

2. (a) Compute $(1+x)^5$ in $\mathbb{Z}_5[x]$.
 (b) Compute $(1+x)^7$ in $\mathbb{Z}_7[x]$.
 (c) Show that $(1+x)^p = 1 + x^p$ in $\mathbb{Z}_p[x]$, if p is a prime. [Hint: Lemma 1 §3.4.]
3. (a) How many polynomials of degree 3 are there in $\mathbb{Z}_5[x]$?
 (b) How many monic polynomials of degree 3 are there in $\mathbb{Z}_3[x]$?
4. (a) Find all roots of $(x-4)(x-5)$ in \mathbb{Z}_6 ; in \mathbb{Z}_7 .
 (b) Find all roots of $x^3 - x$ in \mathbb{Z}_6 ; in \mathbb{Z}_4 .
5. (a) Find the number of roots of $x^2 - x$ in \mathbb{Z}_4 ; $\mathbb{Z}_2 \times \mathbb{Z}_2$; any integral domain; \mathbb{Z}_6 .
 (b) Find a commutative ring in which $x^2 - x$ has infinitely many roots. [Hint: Exercise 29 §3.1; or consider $R = F \times F \times F \times \dots$ where F is a field.]
6. Assume that f, g , and $f + g$ are all nonzero in $R[x]$.
 (a) Show that $\deg(f + g) \leq \max\{\deg f, \deg g\}$ for any ring R .
 (b) Provide an example of where equality fails to hold in (a).
7. (a) Let f and g be nonzero polynomials in $R[x]$ and assume that the leading coefficient of one of them is a unit. Show that $fg \neq 0$ and that $\deg(fg) = \deg f + \deg g$.
 (b) If R is not a domain, show that linear polynomials f and g exist in $R[x]$ such that $\deg(fg) < \deg f + \deg g$.
8. Let R be a subring of S , let $f \neq 0$ and g be polynomials in $R[x]$, and assume that the leading coefficient of f is a unit in R . If f divides g in $S[x]$, show that f divides g in $R[x]$. [Hint: Division algorithm.]
9. Show that $R[x]$ and R have the same characteristic for any ring R .
10. Where is the fact that $a \in Z(R)$ used in the proof of Theorem 5?
11. If a is a nonzero root of $f = a_0 + a_1x + \dots + a_{n-1}x^{n-1} + a_nx^n$, show that a^{-1} (if it exists) is a root of $g = a_n + a_{n-1}x + \dots + a_1x^{n-1} + a_0x^n$. Assume that R is commutative.
12. If a, b , and c are real and $f = x^2 - (a+c)x + (ac - b^2)$, show that every complex root of f is real.
13. Divide $x^3 - 4x + 5$ by $2x + 1$ in $\mathbb{Q}[x]$. Why is it impossible in $\mathbb{Z}[x]$?
14. In each case write $g = qf + r$ in $R[x]$, where $r = 0$ or $\deg r < \deg f$.
 - (a) $g = x^5 + 4x^4 + x^3 + 5x^2 + x + 2$, $f = x^2 + x + 1$, $R = \mathbb{Z}_6$
 - (b) $g = x^5 + 2x^4 + x^2 + x + 4$, $f = x^2 + x + 2$, $R = \mathbb{Z}_7$
 - (c) $g = x^3 + x^2 + 3x + 2$, $f = 3x + 1$, $R = \mathbb{Z}_8$
 - (d) $g = x^3 + 2x^2 + x + 3$, $f = 3x + 2$, $R = \mathbb{Z}_7$
 - (e) $g = 3x^3 + 2x^2 - 8x + 1$, $f = x^2 + 2$, $R = \mathbb{Q}$
 - (f) $g = 3x^3 + 5x^2 + x + 6$, $f = 2x^2 + 1$, $R = \mathbb{Q}$
15. Which of $x - 1$, $x + 1$, and $x - 2$ is a factor of $x^4 - 2x^3 - x^2 + 3x - 2$ in $\mathbb{Z}[x]$?
16. (a) For which primes p is $x - 1$ a factor of $f = 3x^4 + 5x^3 + 2x^2 + x + 4$ in $\mathbb{Z}_p[x]$?
 (b) For which primes p is $x + 2$ a factor of $f = 5x^4 - 2x^3 + 3x^2 + 4x - 1$ in $\mathbb{Z}_p[x]$?
17. In each case factor f into linear factors in $F[x]$.
 - (a) $f = x^4 + 12$, $F = \mathbb{Z}_{13}$
 - (b) $f = x^3 + 1$, $F = \mathbb{Z}_7$
 - (c) $f = x^3 - x^2 + x - 1$, $F = \mathbb{Z}_5$
 - (d) $f = x^3 + 4x^2 + 3x + 5$, $F = \mathbb{Z}_7$
18. Let $a \neq 0$ in a field F . Determine the integers $n \geq 1$ such that $x + a$ is a factor of $x^n + a^n$ in $F[x]$? In these cases write down the factorization.
19. If F is a field, let u, v , and w be distinct roots of $f = x^3 + ax^2 + bx + c$ in $F[x]$. Show that $a = -(u + v + w)$, $b = uv + uw + vw$, and $c = -uvw$.
20. Show that the factor theorem is false in $R[x]$ if R is a noncommutative division ring.
 [Hint: Consider $f = bx - ba$ where $ab \neq ba$ in R .]

21. (a) Show that $\mathbb{Z}_4[x]$ has infinitely many units and infinitely many nilpotents.
 (b) Find a polynomial in $\mathbb{Z}_4[x]$ that is neither a unit nor a nilpotent.
22. If R is a commutative ring, show that the only idempotents in $R[x]$ are in R . [Hint: If $a = 2ae, e^2 = e$, show that $a = 0$.]
23. In each case determine the multiplicity of a as a root of f .
 (a) $f = x^3 - 2x^2 - 4x + 3; a = 3, R = \mathbb{Z}_6$
 (b) $f = x^4 + 2x^2 + 2x + 2; a = -1, R = \mathbb{Z}_3$
 (c) $f = x^5 + 2x^4 + x^3 - x^2 + 2x - 1; a = 1, R = \mathbb{Z}_4$
 (d) $f = 4x^4 - 8x^3 + x^2 - 3x + 9; a = \frac{3}{2}, R = \mathbb{Q}$
24. If R is a commutative ring, a polynomial f in $R[x]$ is said to **annihilate** R if $f(a) = 0$ for every $a \in R$.
 (a) Show that $x^p - x$ annihilates \mathbb{Z}_p . [Hint: Fermat's theorem.]
 (b) Show that $x^5 - x$ annihilates \mathbb{Z}_{10} .
 (c) If $p \neq 2$ is a prime, show that $x^p - x$ annihilates \mathbb{Z}_{2p} . [Hint: Corollary 1, Theorem 8, §3.4.]
 (d) If $p > 3$ is a prime, show that $x^p - x$ annihilates \mathbb{Z}_{3p} . [Hint: As in (c).]
 (e) Does $x^5 - x$ or $x^7 - x$ annihilate \mathbb{Z}_{35} ? Justify your answer.
 (f) Show that a polynomial of degree n exists in $\mathbb{Z}_n[x]$ that annihilates \mathbb{Z}_n .
25. In each case find all rational roots of f and factor f as far as possible in $\mathbb{Q}[x]$.
 (a) $f = 4x^4 + x^3 - 3x^2 + 4x - 3$
 (b) $f = 4x^4 + 4x^3 + 3x^2 - x - 1$
 (c) $f = x^4 - x^3 - x^2 - x - 2$
 (d) $f = x^5 - x^4 + x^3 - x^2 + x - 1$
 (e) $f = x^4 + x^3 + 3x^2 + 2x + 2$
 (f) $f = x^4 - \frac{5}{2}x^3 + \frac{5}{2}x^2 - \frac{5}{2}x + \frac{3}{2}$
26. Show that $\sqrt[m]{m}$ is not rational unless $m = k^n$ for some integer k .
27. If f is a monic polynomial in $\mathbb{Z}[x]$, show that the only rational roots (if any) are integers.
28. If R is an integral domain and $f \in R[x]$ has infinitely many roots in R , show that $f = 0$ is the zero polynomial.
29. Let f and g be polynomials in $R[x]$, where R is an integral domain, and assume that each is either 0 or has degree at most n . If $f(a) = g(a)$ holds for $n+1$ distinct elements $a \in R$, show that $f = g$.
30. Show that $x^p - x = x(x-1)(x-2)\cdots(x-p+1)$ in $\mathbb{Z}_p[x]$, where p is a prime. [Hint: Exercise 24(a) and Theorem 8.]
31. Show that $\langle x \rangle$ is a maximal ideal in $R[x]$ if R is a field. What can be said if R is an integral domain?
32. Let R be any ring and let A denote the set of all polynomials in $R[x]$ whose coefficients sum to 0. Show that A is an ideal of $R[x]$ and $R[x]/A \cong R$ as rings.
33. Define $\theta : R[x] \rightarrow R$ by taking $\theta(f)$ to be 0 or the leading coefficient of f , depending on whether $f = 0$ or $f \neq 0$. Is θ a ring homomorphism? Justify your answer.
34. Let R be a commutative ring and let $\varphi_a : R[x] \rightarrow R$ be evaluation at $a \in R$.
 (a) Show that $\varphi_a(r) = r$ for all $r \in R$.
 (b) If $\theta : R[x] \rightarrow R$ is any ring homomorphism such that $\theta(r) = r$ for all $r \in R$, show that $\theta = \varphi_a$ for some $a \in R$.
 (c) Find a nonzero ring homomorphism $\mathbb{C}[x] \rightarrow \mathbb{C}$ that is not an evaluation.
 (d) Is $a \mapsto \varphi_a$ a ring homomorphism $R \rightarrow F(R[x], R)$? Here $F(R[x], R)$ is the ring of functions $R[x] \rightarrow R$ with pointwise operations. (See Example 4 §3.1.)

35. If A is an ideal of R , let $\mathcal{A}[x] = \{a_0 + r_1x + r_2x^2 + \cdots \mid a_0 \in A, r_i \in R\}$. Show that $\mathcal{A}[x]$ is an ideal of $R[x]$ and $R[x]/\mathcal{A}[x] \cong R/A$.
36. Show that \mathbb{Z}_p can be embedded in an infinite field. [Hint: Theorem 2 and Theorem 5, §3.2.]
37. Let $r \mapsto \bar{r}$ denote a ring homomorphism $\theta : R \rightarrow S$. If $f = a_0 + a_1x + \cdots + a_nx^n$ in $R[x]$, let $\bar{f} = \bar{a}_0 + \bar{a}_1x + \cdots + \bar{a}_nx^n$ in $S[x]$. Prove each of the following statements:
- $\bar{\theta} : R[x] \rightarrow S[x]$ with $\bar{\theta}(f) = \bar{f}$ is a ring homomorphism, onto if θ is onto.
 - If $\ker \theta = A$, then $\ker \bar{\theta} = A[x]$.
 - If $R \cong S$, then $R[x] \cong S[x]$.
 - If A is an ideal of R , then $R[x]/A[x] \cong (R/A)[x]$.
 - If \bar{f} has no root in S , show that f has no root in R .
38. Let R be a commutative ring. Use the notation of Exercise 37.
- If P is a prime ideal of R , show that $P[x]$ is a prime ideal of $R[x]$.
 - If M is a maximal ideal of R , show that $M[x]$ is not a maximal ideal of $R[x]$.
39. Let R be a commutative ring and consider $f = a_0 + a_1x + \cdots + a_nx^n$ in $R[x]$. If a_0 is a unit in R and a_i is nilpotent for all $i \geq 1$, show that f is a unit in $R[x]$. [Hint: If u is a unit and a is nilpotent then $u+a$ is a unit.]
40. If R is commutative and $f \in R[x]$, denote the corresponding polynomial function by $\tilde{f} : R \rightarrow R$. Thus, $\tilde{f}(a) = f(a)$ for every $a \in R$. Let $F(R, R)$ denote the set of all functions $R \rightarrow R$ using pointwise operations (see Example 4 §3.1). Define $\theta : R[x] \rightarrow F(R, R)$ by $\theta(f) = \tilde{f}$.
- Show that θ is a ring homomorphism and hence that the set $P(R, R) = \theta(R[x])$ of all polynomial functions $R \rightarrow R$ is a subring of $F(R, R)$.
 - Show that $\ker \theta = \{f \in R[x] \mid f(a) = 0 \text{ for all } a \in R\}$. These polynomials are said to annihilate R (see Exercise 24).
 - If R is an infinite integral domain, show that $R[x] \cong P(R, R)$.
 - If R is a finite ring, can $R[x] \cong P(R, R)$? Give reasons.
41. **Lagranges Interpolation Expansion.** Let F be a field and let $a_0, a_1, a_2, \dots, a_n$ be distinct elements of F , $n \geq 1$. Define the **Lagrange polynomials**

$$c_k = \frac{\prod_{i \neq k} (x - a_i)}{\prod_{i \neq k} (a_k - a_i)}, \quad k = 0, 1, \dots, n,$$

where the numerator is the product $(x - a_0)(x - a_1) \cdots (x - a_n)$ with $(x - a_k)$ omitted, and the denominator is similar. If $f = 0$ or $\deg f \leq n$ in $F[x]$, show that

$$f = f(a_0)c_0 + f(a_1)c_1 + \cdots + f(a_n)c_n.$$

[Hint: Exercise 29.]

4.2 FACTORIZATION OF POLYNOMIALS OVER A FIELD

The prime factorization theorem (Theorem 7 §1.2) asserts that every integer $n \geq 2$ can be written uniquely as a product of primes. It may come as a surprise that, if F is a field, an analogous factorization theorem holds in the polynomial ring $F[x]$. We devote this section to proving this theorem and to discussing several other results which arise along the way.

The prime integers can be described as follows: An integer $p \geq 2$ is a prime if $p = ab$, where $a, b \in \mathbb{Z}$, implies that either $a = \pm 1$ or $b = \pm 1$. In other words, p

admits only **trivial** factorizations where one of the factors is a unit in \mathbb{Z} . If F is a field, the units of the integral domain $F[x]$ are just the nonzero elements of F (Theorem 2 §4.1). If $a \neq 0$ in F , each polynomial f in $F[x]$ certainly admits the trivial factorization $f = a(a^{-1}f)$. If we rule out such factorizations, we arrive at the following analogue for $F[x]$ of the definition of primes in \mathbb{Z} .

If F is a field, a polynomial $p \neq 0$ in $F[x]$ is called⁵⁸ an **irreducible polynomial** (and we say that p is **irreducible over F**) if:

- (1) $\deg p \geq 1$.
- (2) If $p = fg$ in $F[x]$, then either $\deg f = 0$ or $\deg g = 0$.

Polynomials that are not irreducible are called **reducible**.

Note that $\deg f = 0$ if and only if f is a nonzero constant in $F[x]$, that is f is a unit in $F[x]$ by Theorem 2 §4.1. Hence, condition (1) ensures that no irreducible polynomial is a unit (the analogous condition on \mathbb{Z} is that $p \geq 2$ for all primes p).

Example 1. Show that every linear polynomial is irreducible over any field F .

Solution. Let $p = ax + b$, $a \neq 0$, be linear, and suppose that $p = fg$ in $F[x]$. As F is a domain, Theorem 2 §4.1 gives $\deg f + \deg g = \deg p = 1$. As both $\deg f \geq 0$ and $\deg g \geq 0$ are integers, one of them must equal 0. \square

Example 2. If F is a field and p is irreducible in $F[x]$, show that ap is also irreducible for all $a \neq 0$ in F .

Solution. If $ap = fg$ in $F[x]$, then $p = (a^{-1}f)g$. Because p is irreducible, it follows that either $\deg g = 0$ or $\deg(a^{-1}f) = 0$. As $\deg f = \deg(a^{-1}f)$, ap is irreducible. \square

If f is any irreducible polynomial in $F[x]$, with leading coefficient a , we have $f = ap$, where $p = a^{-1}f$ is irreducible by Example 2 and also monic (leading coefficient 1). Thus, there is no great loss in generality in working with monic irreducible polynomials.

Every linear polynomial in $F[x]$ has the form $p = ax + b$, $a \neq 0$, and so has a root $-a^{-1}b$ in F . However, no irreducible polynomial of degree 2 or more can have a root in F .

Theorem 1. Let F be a field and consider p in $F[x]$ where $\deg p \geq 2$.

- (1) If p is irreducible, then p has no root in F .
- (2) If $\deg p$ is 2 or 3 then p is irreducible if and only if it has no root in F .

Proof. (1) If p has a root $a \in F$, then $p(a) = 0$, so $p = (x - a)q$ by the factor theorem. Because $\deg p \geq 2$, this means p is not irreducible, contrary to hypothesis. Hence p has no root in F .

(2) Assume that p has no root in F . If $p = fg$ then f and g have no root in F , so $\deg f \neq 1$ and $\deg g \neq 1$. But $\deg f + \deg g = \deg p$ is 2 or 3, so $\deg f = 0$ or $\deg g = 0$. Hence p is irreducible. The converse follows from (1). \blacksquare

⁵⁸They are not called *prime* polynomials; that term is reserved for the stronger property: If $p|fg$ then $p|f$ or $p|g$. This will be fully investigated in Section 5.1.

Example 3. $x^2 + 1$ is irreducible in $\mathbb{R}[x]$ because it has no root in \mathbb{R} .

Example 4. Determine if $p = x^3 + 3x^2 + x + 2$ is irreducible over \mathbb{Z}_5 .

Solution. Because $\mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$, we can compute $p(0) = 2$, $p(1) = 2$, $p(2) = 4$, $p(3) = 4$, and $p(4) = 3$. Hence p has no root in \mathbb{Z}_5 and so is irreducible. \square

Example 5. Show that $x^2 + x + 1$ is the only irreducible quadratic over \mathbb{Z}_2 .

Solution. Because $\mathbb{Z}_2 = \{0, 1\}$, every quadratic in $\mathbb{Z}_2[x]$ has the form $x^2 + ax + b$, where a and b lie in \mathbb{Z}_2 . Hence there are four possibilities: x^2 , $x^2 + x$, $x^2 + 1$, and $x^2 + x + 1$. The first three have a root in \mathbb{Z}_2 , whereas $x^2 + x + 1$ does not. Hence $x^2 + x + 1$ is the only irreducible quadratic in $\mathbb{Z}_2[x]$. \square

Part (2) of Theorem 1 provides a useful test of irreducibility for polynomials of degree 2 or 3, but it fails for polynomials of degree 4 or more. For example,

$$p = x^4 + 3x^2 + 2 = (x^2 + 1)(x^2 + 2)$$

is clearly not irreducible in $\mathbb{R}[x]$, but it has no root in \mathbb{R} .

Example 6. Show that $p = x^2 - 2$ is irreducible over \mathbb{Q} but not over \mathbb{R} .

Solution. $p = (x - \sqrt{2})(x + \sqrt{2})$ in $\mathbb{R}[x]$, so evidently it is not irreducible over \mathbb{R} . But this expression shows that the only roots of p in \mathbb{R} are $\sqrt{2}$ and $-\sqrt{2}$, and neither is in \mathbb{Q} . Hence p is irreducible over \mathbb{Q} by Theorem 1. \square

Observe that Example 6 shows that the phrase “ p is irreducible” is meaningless unless we specify which field is to be used for the coefficients.

Example 7. If $p \equiv 3 \pmod{4}$, p a prime, show that $x^2 + 1$ is irreducible over \mathbb{Z}_p .

Solution. If $a \neq 0$ in \mathbb{Z}_p , Fermat’s theorem (Theorem 8 §1.3) gives $a^{p-1} = 1$. If $p = 4k + 3$, $k \geq 0$, then $1 = a^{p-1} = a^{4k+2} = (a^2)^{2k+1}$. Hence, $a^2 = -1$ is impossible in \mathbb{Z}_p , so $x^2 - 1$ has no root in \mathbb{Z}_p . \square

Note that the converse of Example 7 is also true (Exercise 17); that is, $x^2 + 1$ is irreducible over \mathbb{Z}_p (p a prime) if and only if $p \equiv 3 \pmod{4}$.

If F is a field, the irreducible polynomials are the analogues in $F[x]$ of the primes in \mathbb{Z} . Since there is no way to systematically write down all the integral primes, any characterization of all the irreducible polynomials in $F[x]$ would seem difficult (which, in fact, is the case). However, an explicit description does exist in the case of $F = \mathbb{C}$ or $F = \mathbb{R}$. This depends on a deep theorem first proved in 1799 by Gauss.

Theorem 2. Fundamental Theorem of Algebra. If $f \in \mathbb{C}[x]$ is a nonconstant polynomial, then f has a root in \mathbb{C} .

This result is sometimes express by saying that the complex field is *algebraically closed*, and we have more to say about that in Chapter 6. No simple proofs of the theorem are known, and most proofs involve analysis at some stage. We give one proof in Section 6.6.

Theorem 3. As usual, let \mathbb{C} denote the field of complex numbers.

- (1) If $\deg f = n \geq 1$, $f \in \mathbb{C}[x]$, then f factors completely as

$$f = u(x - u_1)(x - u_2) \cdots (x - u_n)$$

where u is the leading coefficient of f and u_1, u_2, \dots, u_n are the (not necessarily distinct) roots of f in \mathbb{C} .

- (2) The only irreducible polynomials in $\mathbb{C}[x]$ are linear.

Proof. (2) is clear by (1). To prove (1) induct on $n = \deg f$. If $n = 1$, then $f = ux + b = u(x + u^{-1}b)$, so $u_1 = -u^{-1}b$. If $n > 1$, then f has a root u_1 by Theorem 2, so $f = (x - u_1)q$ where $\deg q = n - 1$. Then $q = u(x - u_2) \cdots (x - u_n)$ by induction, so $f = u(x - u_1)(x - u_2) \cdots (x - u_n)$ has the desired form. Clearly, u is the leading coefficient of f , and u_1, u_2, \dots, u_n are the roots. ■

The fundamental theorem shows the existence of roots of complex polynomials but reveals no way to find them. This is difficult in general. Even so, the theorem has many applications, as illustrated by the analysis it provides of real polynomials.

Let $q = ax^2 + bx + c$, $a \neq 0$, be a real quadratic. If u is a root of q in \mathbb{C} , we have $au^2 + bu + c = 0$. We solve for u by using the famous **quadratic formula**:

$$u = \frac{1}{2a} \left[-b \pm \sqrt{b^2 - 4ac} \right].$$

The quantity $b^2 - 4ac$ is called the **discriminant** of q . If q is irreducible, it has no real roots so $b^2 < 4ac$ and the two nonreal complex roots are conjugates:

$$u = \frac{1}{2a} \left[-b + i\sqrt{4ac - b^2} \right] \quad \text{and} \quad \bar{u} = \frac{1}{2a} \left[-b - i\sqrt{4ac - b^2} \right].$$

The converse is also true: If u is any nonreal complex number, then u and \bar{u} are the roots of an irreducible real quadratic. In fact, the (monic) quadratic

$$x^2 - (u + \bar{u})x + u\bar{u} = (x - u)(x - \bar{u})$$

has real coefficients $u\bar{u} = |u|^2$ and $u + \bar{u} = 2 \operatorname{re} u$, and so is irreducible over \mathbb{R} because its roots u and \bar{u} are not real.

Theorem 4. Every nonconstant polynomial f in $\mathbb{R}[x]$ factors as

$$f = a(x - r_1)(x - r_2) \cdots (x - r_m) q_1 q_2 \cdots q_k,$$

where a is the leading coefficient of f ; r_1, r_2, \dots, r_m are the real roots of f (if any); and q_1, q_2, \dots, q_k are monic irreducible real quadratics (perhaps none).

Proof. Write $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$, where the coefficients a_i are real. If u is a complex root of f , we claim first that the conjugate \bar{u} is also a root. Indeed, $f(u) = 0$, so

$$\begin{aligned} 0 &= \bar{0} = \overline{f(u)} = \overline{a_0 + a_1u + a_2u^2 + \cdots + a_nu^n} \\ &= \bar{a}_0 + \bar{a}_1\bar{u} + \bar{a}_2\bar{u}^2 + \cdots + \bar{a}_n\bar{u}^n \\ &= a_0 + a_1\bar{u} + a_2\bar{u}^2 + \cdots + a_n\bar{u}^n \\ &= f(\bar{u}), \end{aligned}$$

where $\bar{a}_i = a_i$ for each i because a_i is real. Thus, the nonreal roots of f (if any) come in conjugate pairs u and \bar{u} . Hence, f factors in $\mathbb{C}[x]$ as

$$f = a(x - r_1)(x - r_2) \cdots (x - r_m)(x - u_1)(x - \bar{u}_1) \cdots (x - u_k)(x - \bar{u}_k)$$

by Theorem 3, where $a = a_n$ is the leading coefficient of f ; r_1, \dots, r_m are the real roots; and $u_1, \bar{u}_1, \dots, u_m, \bar{u}_m$ are the pairs of nonreal roots. This proves the

theorem because each product

$$q_j = (x - u_j)(x - \bar{u}_j) = x^2 - (u_j + \bar{u}_j)x + u_j\bar{u}_j$$

is an irreducible real quadratic (see the discussion preceding this theorem). ■

As an immediate consequence of Theorem 4, we have

Corollary. *The irreducible polynomials in $\mathbb{R}[x]$ are either linear or quadratic.*

Irreducibles Over the Rationals

Theorems 3 and 4 completely describe the irreducible polynomials in $\mathbb{C}[x]$ and $\mathbb{R}[x]$. However, the situation in $\mathbb{Q}[x]$ is much more complicated. If $f \in \mathbb{Q}[x]$ and m is the least common multiple of the denominators of the coefficients of f , then $mf \in \mathbb{Z}[x]$. So it is not surprising that many questions about $\mathbb{Q}[x]$ come down to questions about $\mathbb{Z}[x]$. Theorem 5 is the key to making this transition from $\mathbb{Q}[x]$ to $\mathbb{Z}[x]$.

Theorem 5. Gauss' Lemma. *Let $f = gh$ in $\mathbb{Z}[x]$. If a prime $p \in \mathbb{Z}$ divides every coefficient of f , then either p divides every coefficient of g or p divides every coefficient of h .*

Before giving the proof we introduce an important homomorphism. Given a prime $p \in \mathbb{Z}$ and an integer a , we let \bar{a} denote the residue of a in \mathbb{Z}_p . If we are given $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ in $\mathbb{Z}[x]$, the polynomial

$$\bar{f} = \bar{a}_0 + \bar{a}_1x + \bar{a}_2x^2 + \cdots + \bar{a}_nx^n \quad \text{in } \mathbb{Z}_p[x]$$

is called the **reduction** of f modulo p . The point is that $f \mapsto \bar{f}$ is an onto ring homomorphism $\mathbb{Z}[x] \rightarrow \mathbb{Z}_p[x]$ —see Exercise 37 §4.1.

Proof of Gauss' Lemma. Let $f = gh$ in $\mathbb{Z}[x]$ and suppose that p divides every coefficient of f . Then $\bar{f} = 0$ in $\mathbb{Z}_p[x]$ so, as reduction modulo p is a homomorphism, we get $0 = \bar{f} = \bar{g} \cdot \bar{h}$. But $\mathbb{Z}_p[x]$ is an integral domain (\mathbb{Z}_p is a field), so this means $\bar{g} = 0$ or $\bar{h} = 0$. But $\bar{g} = 0$ in $\mathbb{Z}_p[x]$ means that every coefficient is zero in \mathbb{Z}_p , that is, every coefficient of g is divisible by p . Gauss' lemma follows. ■

If f is a nonconstant polynomial in $\mathbb{Z}[x]$, we say that $f = gh$ in $\mathbb{Z}[x]$ is a **proper factorization** if both g and h have smaller degree than f . Then Gauss' lemma yields the following useful theorem.

Theorem 6. *Let f be a nonconstant polynomial in $\mathbb{Z}[x]$.*

- (1) *If $f = gh$ with g and h in $\mathbb{Q}[x]$, then $f = g_0h_0$ where g_0 and h_0 are in $\mathbb{Z}[x]$, $\deg g_0 = \deg g$, and $\deg h_0 = \deg h$.*
- (2) *f is irreducible in $\mathbb{Q}[x]$ if and only if it has no proper factorization in $\mathbb{Z}[x]$.*

Proof. (1) Let a and b be the least common multiples of the denominators of the coefficients of g and h , respectively. Then $g_1 = ag$ and $h_1 = bh$ are in $\mathbb{Z}[x]$, so

$$abf = g_1h_1$$

is an equation in $\mathbb{Z}[x]$. If p is a prime dividing ab , then Gauss' lemma shows that either p divides all coefficients of g_1 or p divides all coefficients of h_1 . Hence, p can

be canceled to give a new equation in $\mathbb{Z}[x]$:

$$\frac{ab}{p} f = g_2 h_2.$$

Continuing, we delete every prime factor of ab , and finally obtain a factorization

$$f = g_k h_k$$

in $\mathbb{Z}[x]$. Now (1) follows because each of the polynomials g_1, g_2, \dots has the same degree as g and, similarly, $\deg h_i = \deg h$ for each i .

(2) If f is irreducible in $\mathbb{Q}[x]$, it has no proper factorization in $\mathbb{Q}[x]$ and hence none in $\mathbb{Z}[x]$. The converse follows from (1). \blacksquare

Incidentally, a polynomial in $\mathbb{Z}[x]$ that has no proper factorization in $\mathbb{Z}[x]$ is not called “irreducible in $\mathbb{Z}[x]$ ”. The reason is that, in the general factorization theory to be developed in Chapter 5, polynomials such as $5x - 5 = 5(x - 1)$ are not called irreducible in $\mathbb{Z}[x]$ even though they admit no proper factorization in $\mathbb{Z}[x]$.

Given f in $\mathbb{Q}[x]$, there is an integer $a \neq 0$ such that af is in $\mathbb{Z}[x]$ (any common multiple a of the denominators of the nonzero coefficients of f will do). By Example 2, f is irreducible in $\mathbb{Q}[x]$ if and only if af is irreducible in $\mathbb{Q}[x]$. Hence, Theorem 6 reduces the problem of testing whether f is irreducible in $\mathbb{Q}[x]$ to the problem of showing that af has no proper factorization in $\mathbb{Z}[x]$. So we look at this latter problem. The following observation is relevant:

Lemma 1. *If $f \in \mathbb{Z}[x]$ is monic and $f = gh$ in $\mathbb{Z}[x]$, we may assume that both g and h are monic.*

Proof. Let the leading coefficients of g and h be a and b , respectively, so the leading coefficient of gh is ab . Hence, the fact that $f = gh$ is monic means $1 = ab$ so, since $a, b \in \mathbb{Z}$, either $a = b = 1$ or $a = b = -1$. Now Lemma 1 follows. \blacksquare

Example 8. Show that $f = x^5 + 2x^2 + 1$ is irreducible in $\mathbb{Q}[x]$.

Solution. By Theorem 6 we show that f has no proper factorization in $\mathbb{Z}[x]$. It has no linear factors because it has no rational roots (the only candidates are ± 1 by the rational roots theorem). Hence, if f factors at all in $\mathbb{Z}[x]$, it factors as a quadratic times a cubic. Moreover, the factors may be taken to be monic by Lemma 1 because f is monic, say

$$f = x^5 + 2x^2 + 1 = (x^2 + ax + b)(x^3 + cx^2 + dx + e),$$

where a, b, c, d , and e are integers. Multiplying the right side out and equating coefficients of powers of x gives five equations:

$$a + c = 0, \quad d + ac + b = 0, \quad e + ad + bc = 2, \quad ae + bd = 0, \quad \text{and} \quad be = 1.$$

The last equation gives $b = e = \pm 1$; then the second-to-last equation gives $a + d = 0$. Because $a + c = 0$ too, the second equation becomes $-a - a^2 + b = 0$. Hence a is an integral root of $x^2 + x - b = 0$, where $b = \pm 1$. But the only rational candidates are ± 1 by the rational roots theorem, and neither is a root when b is 1 or -1 . Hence no such factorization of f exists, so f is irreducible in $\mathbb{Q}[x]$. \square

The heavy-handed method in Example 8 is less effective for polynomials f of higher degree, because the resulting systems of equations are complicated.

Therefore, we give two other irreducibility tests for $\mathbb{Q}[x]$. The first utilizes, for a prime p , the homomorphism $\mathbb{Z}[x] \rightarrow \mathbb{Z}_p[x]$ used in the proof of Gauss' lemma, where $g \mapsto \bar{g}$ and \bar{g} comes from reducing coefficients of g modulo p .

Theorem 7. Modular Irreducibility Test. Let $0 \neq f \in \mathbb{Z}[x]$ and suppose that a prime p exists such that

- (1) p does not divide the leading coefficient of f —for example if f is monic.
- (2) The reduction \bar{f} of f modulo p is irreducible in $\mathbb{Z}_p[x]$.

Then f is irreducible in $\mathbb{Q}[x]$.

Proof. First, $\deg \bar{f} = \deg f$ by condition (1). Suppose f is not irreducible in $\mathbb{Q}[x]$, so there is a proper factorization $f = gh$ in $\mathbb{Z}[x]$ by Theorem 6. Then, we have $\deg \bar{g} \leq \deg g < \deg f = \deg \bar{f}$ and, similarly, $\deg \bar{h} < \deg \bar{f}$. But $\bar{f} = \bar{g}\bar{h}$, contradicting the irreducibility of \bar{f} in $\mathbb{Z}_p[x]$. ■

Example 9. Show that $f = x^3 + 4x^2 + 6x + 2$ is irreducible in $\mathbb{Q}[x]$.

Solution. We could use the rational roots theorem to show that f has no root in \mathbb{Q} . However, reducing modulo 3 is much easier. Then $\bar{f} = x^3 + x^2 + 2$, which clearly has no root in \mathbb{Z}_3 . Hence Theorem 7 applies. □

Example 10. Show that $f = x^4 + 2x^3 + 2x^2 - x + 1$ is irreducible in $\mathbb{Q}[x]$.

Solution. Reduction modulo 2 gives $\bar{f} = x^4 + x + 1$ in $\mathbb{Z}_2[x]$. This polynomial has no root in \mathbb{Z}_2 so, if it fails to be irreducible, it must factor into two quadratics. These must be irreducible (they have no root) so, by Example 5, both must equal $x^2 + x + 1$. But $(x^2 + x + 1)^2 = x^4 + x^2 + 1 \neq \bar{f}$. Hence \bar{f} is irreducible in $\mathbb{Z}_2[x]$, so f is irreducible in $\mathbb{Q}[x]$ by Theorem 7. □

Note that the converse of the modular irreducibility test fails. In fact $x^4 + 1$ is irreducible in $\mathbb{Q}[x]$ but not in $\mathbb{Z}_p[x]$ for *any* prime p (see Example 4 §6.4).

The next test for \mathbb{Q} -irreducibility is due to F.G. Eisenstein, a pupil of Gauss.

Theorem 8. Eisenstein Criterion. Consider $f = a_0 + a_1x + \cdots + a_nx^n$ in $\mathbb{Z}[x]$, where $n \geq 1$ and $a_n \neq 0$. Suppose that a prime $p \in \mathbb{Z}$ exists such that

- (1) p divides each of a_0, a_1, \dots, a_{n-1} .
- (2) p does not divide a_n .
- (3) p^2 does not divide a_0 .

Then f is irreducible in $\mathbb{Q}[x]$.

Proof. If it is not irreducible, let $f = gh$ be a proper factorization in $\mathbb{Z}[x]$ (by Theorem 6). Write

$$g = b_0 + b_1x + \cdots + b_mx^m \quad \text{and} \quad h = c_0 + c_1x + \cdots + c_tx^t.$$

Because p divides $a_0 = b_0c_0$ and p^2 does not divide a_0 , it follows that p divides exactly one of b_0 or c_0 , say p divides b_0 but not c_0 . Then p does not divide b_m (by (2), as $a_n = b_mc_t$), so let b_k be the first integer in the list b_0, b_1, \dots, b_m not divisible by p . Equating coefficients of x^k in $f = gh$ gives

$$a_k = b_kc_0 + b_{k-1}c_1 + \cdots + b_0c_k.$$

Now p divides a_k (by (1), because $k \leq m < n$), and p divides every term in the sum after the first (by the choice of b_k). Hence p divides $b_k c_0$ too, so it divides one of b_k and c_0 . This contradiction proves the theorem. ■

Example 11. Show that $2x^5 + 27x^3 - 18x + 12$ is irreducible in $\mathbb{Q}[x]$.

Solution. The Eisenstein criterion applies with $p = 3$. □

Example 12. $\mathbb{Q}[x]$ contains an irreducible polynomial of every positive degree. In fact $x^n - 2$ is irreducible in $\mathbb{Q}[x]$ for any $n \geq 1$ by the Eisenstein criterion. □

Example 13. If p is a prime, show that the p th **cyclotomic polynomial**

$$\Phi_p = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible in $\mathbb{Q}[x]$.

Solution. Replacing x by $x + 1$, it suffices to show that $\Phi_p(x + 1)$ is irreducible. [Indeed, if $\Phi_p = gh$ is a proper factorization, then the same is true of the factorization $\Phi_p(x + 1) = g(x + 1)h(x + 1)$.] Now observe that

$$(x - 1)\Phi_p = (x - 1)(x^{p-1} + x^{p-2} + \cdots + x + 1) = x^p - 1.$$

Replacing x by $x + 1$ gives $x\Phi_p(x + 1) = (x + 1)^p - 1$ so, by the binomial theorem,

$$\Phi_p(x + 1) = x^{p-1} + \binom{p}{1}x^{p-2} + \cdots + \binom{p}{p-2}x + p.$$

But p divides $\binom{p}{k}$ for $1 \leq k \leq p - 2$ by Lemma 1 §3.4. Hence, the Eisenstein criterion applies, showing that $\Phi_p(x + 1)$ is irreducible. □

Unique Factorization

Theorems 3 and 4 show that any polynomial in $\mathbb{C}[x]$ or $\mathbb{R}[x]$ is a constant times a product of (monic) irreducible factors. We conclude this section with a proof that this is true in $F[x]$ for any field F , and that the resulting factorization is unique.

One comment on uniqueness is in order. The prime factorization theorem for \mathbb{Z} asserts that every integer $n \geq 2$ is uniquely a product of primes. However, the uniqueness requires the assumption that primes are positive. Hence every integer apart from 0, 1, and -1 factors uniquely as a unit ± 1 times a product of primes. The exceptions are 0 and the units ± 1 of \mathbb{Z} . If F is a field, the units in $F[x]$ are the nonzero constant polynomials, so the analogue for $F[x]$ of the prime factorization theorem would be: Every nonconstant polynomial in $F[x]$ factors uniquely as a unit $u \neq 0$ in F times a product of irreducible polynomials. But because of the trivial factorization $f = a(a^{-1}f)$, uniqueness here requires that the irreducible polynomials be monic (this is analogous to insisting that primes in \mathbb{Z} are positive). The reason this works is Theorem 9.

Recall that, if R is commutative ring, we say that a polynomial $d \in R[x]$ is a **divisor** of $f \in R[x]$, or that d **divides** f , if $f = qd$ for some $q \in F[x]$.

Theorem 9. Let F be a field and let f and g be nonzero monic polynomials in $F[x]$, each of which divides the other. Then $f = g$.

Proof. If $f = qg$ and $g = pf$ in $R[x]$ then eliminating g gives $f = qp f$. Hence $1 = qp$ ($F[x]$ is a domain) so q is a constant in F . Since $f = qg$ and f and g are monic, comparing leading coefficients gives $q = 1$. Hence $f = g$. ■

The proof of the following result is left to the reader (Exercise 40).

Corollary. If F is a field and $p \in F[x]$ is monic, the following are equivalent:

- (1) p is irreducible.
- (2) If d is a monic divisor of p then either $d = 1$ or $d = p$.

With these results, the factorization theory for $F[x]$ closely parallels that for \mathbb{Z} . Therefore, we skip many details and merely sketch the proofs of the results. The first item on the agenda is the notion of greatest common divisor.

Theorem 10. Let f and g be nonzero polynomials in $F[x]$, where F is a field. Then a uniquely determined polynomial d exists in $F[x]$ satisfying the following conditions.

- (1) d is monic.
- (2) d divides both f and g .
- (3) If h divides both f and g , then h divides d .
- (4) $d = uf + vg$ for some polynomials u and v in $F[x]$.

Finally, d is the unique polynomial satisfying (1), (2), and (3).

Proof. Consider $X = \{uf + vg \mid u, v \text{ in } F[x]\}$. This set contains nonzero polynomials (for example f^2) and thus contains monic polynomials. Among all the monic polynomials in X , let $d = uf + vg$ be one of smallest degree. Then (1) and (4) are satisfied, and (3) is an easy consequence of (4). By the division algorithm write $f = qd + r$, where $r = 0$ or $\deg r < \deg d$. Then

$$r = f - qd = f - q(uf + vg) = (1 - qu)f - (qv)g.$$

If $r \neq 0$ and a is the leading coefficient of r , this expression shows that $a^{-1}r$ is a monic member of X of smaller degree than d . This result contradicts the choice of d , so $r = 0$ and d divides f . Similarly, d divides g , proving (2).

Finally, if d_1 is another polynomial satisfying (1), (2), and (3), then d and d_1 each divides the other. Hence $d = d_1$ by the Corollary to Theorem 9, proving uniqueness. \blacksquare

As in \mathbb{Z} , the polynomial d in Theorem 10 is called the **greatest common divisor** of f and g in $F[x]$, denoted $\gcd(f, g)$, and f and g are called **relatively prime** in $F[x]$ if $d = 1$. Note that 1 is the unique monic polynomial of degree 0, and that Theorem 10 allows the possibility that $d = 1$.

Example 14. Find the greatest common divisor d of $x^2 - 1$ and $2x + 1$ in $\mathbb{Q}[x]$.

Solution. Because d divides the irreducible polynomial $2x + 1$, either $d = 1$ or $d = x + \frac{1}{2}$. But $x + \frac{1}{2}$ does not divide $x^2 - 1$, so $d = 1$. Moreover, the division algorithm gives

$$x^2 - 1 = \left[\frac{1}{4}(2x - 1)\right](2x + 1) - \frac{3}{4}.$$

This implies that $1 = \frac{1}{3}(2x - 1)(2x + 1) - \frac{4}{3}(x^2 - 1)$, and so expresses d as a linear combination of $x^2 - 1$ and $2x + 1$. \square

In general, if $d = \gcd(f, g)$ the analogue of the Euclidean algorithm (Section 1.2) expresses d as a linear combination of f and g . Example 15 provides an illustration.

Example 15. Find the gcd of $f = x^4 - x^2 + x - 1$ and $g = x^3 - x^2 + x - 1$ in $\mathbb{Q}[x]$ and express it as a linear combination of these polynomials.

Solution. We use the division algorithm repeatedly to obtain

$$\begin{aligned} f &= (x+1)g + (-x^2 + x) \\ g &= (-x)(-x^2 + x) + (x - 1) \\ -x^2 + x &= (-x)(x - 1) + 0. \end{aligned}$$

As in \mathbb{Z} , the last nonzero remainder $d = x - 1$ is the gcd. (In this case it happens to be monic; in general, if the leading coefficient is a then d is obtained by multiplying by a^{-1} .) Eliminating remainders gives the required linear combination:

$$\begin{aligned} x - 1 &= g + x(-x^2 + x) = g + x[f - (x+1)g] \\ &= xf - (x^2 + x - 1)g. \end{aligned} \quad \square$$

Theorem 11 is the analogue for polynomials of Euclid's lemma for integers.

Theorem 11. Let $p \in F[x]$ be irreducible, F a field. If p divides a product $f_1 f_2 \cdots f_n$ of nonzero polynomials in $F[x]$, then p divides one of the f_i .

Proof. By induction on n , it suffices to do the case $n = 2$. If p divides fg , let $d = \gcd(f, p)$. Then d divides p so, as p is irreducible, either $\deg d = 0$ (so $d = 1$) or $\deg d = \deg p$. In the second case, $p = ad$, $a \in F$, so p divides f (because d divides f), and we are finished. If $d = 1$, Theorem 10 gives $1 = up + vf$, where $u, v \in F[x]$. Hence $g = ugp + vfg$, so p divides g (because p divides fg). Hence, the theorem holds in this case too. ■

Theorem 12. Unique Factorization Theorem. If F is a field, let f be a nonconstant polynomial in $F[x]$. Then

- (1) $f = ap_1 p_2 \cdots p_m$, where $a \in F$ and p_i is monic and irreducible for all i .
- (2) The factorization in (1) is unique except for the order of the factors.

Proof. (1) It suffices to write f as a product $f = q_1 q_2 \cdots q_m$, where each q_i is irreducible. (If a_i is the leading coefficient of q_i for each i , take $p_i = a_i^{-1} q_i$ and $a = a_1 a_2 \cdots a_m$.) Proceed by strong induction on $n = \deg f$. If $n = 1$, then f itself is irreducible. If $n > 1$, then either f is irreducible (and we're done) or $f = gh$, where $0 < \deg g < n$ and $0 < \deg h < n$. In this case, both g and h are products of irreducible polynomials by induction.

(2) If it is not unique, let f be a nonconstant polynomial of minimal degree that admits two such factorizations:

$$f = ap_1 p_2 \cdots p_m = bq_1 q_2 \cdots q_k.$$

Then $a = b$ because each is the leading coefficient of f . Now Theorem 11 asserts that p_1 divides one of the q_j , say p_1 divides q_1 . Because $\deg p_1 \neq 0$, this implies that $\deg p_1 = \deg q_1$ and hence that $q_1 = cp_1$, $c \in F$. But q_1 and p_1 are monic, so $c = 1$ and $p_1 = q_1$. Canceling gives another polynomial $p_2 \cdots p_m = q_2 \cdots q_k$ of lower degree than f that has two such factorizations. This result contradicts the choice of f . ■

Example 16. Factor $f = x^3 - 1$ into irreducibles in $\mathbb{C}[x]$, $\mathbb{R}[x]$, $\mathbb{Q}[x]$, $\mathbb{Z}_5[x]$, and $\mathbb{Z}_7[x]$.

Solution. We have $f = (x - 1)(x^2 + x + 1)$ over any field. Now $p = x^2 + x + 1$ has no root in \mathbb{R} , \mathbb{Q} , or \mathbb{Z}_5 , so the factorization is $f = (x - 1)p$ in these cases. However, $p = (x - u)(x - \bar{u})$ in $\mathbb{C}[x]$, where $u = \frac{1}{2}(-1 + i\sqrt{3})$, and $p = (x - 2)(x - 4)$ in $\mathbb{Z}_7[x]$. Thus f factors completely into linear factors over \mathbb{C} and \mathbb{Z}_7 . \square

Carl Friedrich Gauss (1777–1855) There is little doubt that Gauss ranks with Archimedes and Newton as one of the greatest mathematicians of all time. He began as a child prodigy and became possibly the last mathematician to know everything in his subject. By the time he was 20 he had, among other things, shown that a polygon of 17 sides was constructible with compass and straightedge (a problem unsolved since the time of the ancient Greeks), discovered the method of least squares (10 years before Legendre), proved that every positive integer is the sum of three triangular numbers (of the form $\frac{1}{2}n(n + 1)$), and proved the law of quadratic reciprocity (a feat that had eluded Euler). At the age of 22 he completed his Ph.D. dissertation under Pfaff at the University of Helmsted by giving the first rigorous proof of the fundamental theorem of algebra. In 1801, he published a timeless masterpiece, *Disquisitiones Arithmeticae*, on number theory in which he introduced the idea of congruence and which made him famous at the age of 24.

Gauss was also gifted in areas other than mathematics. He was very good at languages, and before he was 19 he had seriously considered philology as a profession. (At age 62 he started learning Russian, and in two years was completely literate.) He also had other scientific interests. His discovery of the method of least squares led him to the bell-shaped normal curve in statistics, now called the gaussian distribution. His interests in physics were both theoretical and experimental. He did fundamental work in the theory of electromagnetism (the unit of magnetic intensity is called the Gauss) and, among other things, he invented the electric telegraph with Wilhelm Weber.

Indeed, he is regarded as one of the great physicists. Moreover, astronomers also consider him as one of their own. He spent nearly 40 years as director of the observatory at Göttingen, and when Ceres was discovered and then lost to view, Gauss applied his prodigious computational skill to compute the orbit from the limited data available. The methods he devised are still in use, and Ceres was "rediscovered" precisely where Gauss predicted.

The motto on Gauss's seal was *pauca sed matura*—few but ripe. He lived by this dictum in the sense that he refused to publish any work until he had perfected it. "A cathedral is not a cathedral" he said, "until the last scaffolding is down and out of sight." This led him to withhold publication of several major discoveries because he had not had time to polish them. He wrote them instead in his diary, which ultimately contained 46 cryptic statements of results in 19 pages. The diary was misplaced after his death but reappeared in 1898 and was published (by Felix Klein) in 1901, 46 years after Gauss died. Although not all his results were recorded in the diary (many were set down only in letters to friends), several entries would have each given fame to their author if published. Gauss knew about the quaternions before Hamilton; he invented noneuclidean geometry before Bolyai and Lobachevski; he studied elliptic functions before Abel and Legendre; and, before Cauchy, he had defined analytic functions of a complex variable and proved what is now called the Cauchy integral theorem.

Gauss disliked teaching and preferred his job at the observatory to a professorship. He usually rejected aspiring young mathematicians who approached him; but the students

that he did accept included Eisenstein, Riemann, Kummer, Dirichlet, and Dedekind. His mathematical interests knew no bounds, and many of his achievements have not been mentioned here (his fundamental work in differential geometry, for example, or his apparent possession of the prime number theorem). It is no wonder he is called "the prince of mathematicians." It is nearly 160 years since his death, but, as E. T. Bell has said, "he lives everywhere in mathematics."

Exercises 4.2

1. (a) If $a \neq 0$ in a field F , show that a divides f for every f in $F[x]$.
 (b) If p divides f for every f in $F[x]$, show that $p = a \neq 0$, $a \in F$.
2. If f and g are in $F[x]$, F a field, consider the statements:
 (1) $f = ag$ for $0 \neq a \in F$; (2) f and g have the same roots in F .
 (a) Show that (1) \Rightarrow (2). (b) Does (2) \Rightarrow (1)? Support your answer.
3. In each case explain why f is not irreducible over any field.
 (a) $f = x^3 - 2x^2 + 3x - 2$ (b) $f = x^3 + x^2 + 4$
4. In each case determine whether the polynomial is irreducible. Give reasons.
 (a) $x^3 + 5$ in $\mathbb{Z}_7[x]$ (b) $x^2 - 2$ in $\mathbb{R}[x]$
 (c) $x^2 + 11$ in $\mathbb{C}[x]$ (d) $x^3 - 4$ in $\mathbb{Z}_{11}[x]$
 (e) $x^3 + x + 1$ in $\mathbb{Z}_5[x]$ (f) $x^2 + x + 1$ in $\mathbb{Z}_{17}[x]$
5. In each case determine whether the polynomial is irreducible over each of the fields $\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_5$, and \mathbb{Z}_7 .
 (a) $x^2 - 3$ (b) $x^2 + x + 1$ (c) $x^3 + x + 1$ (d) $x^3 - 2$
6. Let R be an integral domain and let $f \in R[x]$ be monic. If f factors properly in $R[x]$, show that it has a proper factorization $f = gh$, where g and h are both monic.
7. Find a monic quartic in $\mathbb{R}[x]$ with $(1-i)$ and i as roots. Is there a cubic?
8. (a) If $x^2 + ax + b$ has roots u and v in a field F , show that $b = uv$ and $a = -(u+v)$.
 (b) Show that $1+i$ is a root of $x^2 + (1-2i)x - (3+i) \in \mathbb{C}[x]$. Find the other root.
9. Show that an odd degree polynomial in $\mathbb{R}[x]$ has a real root. (Requires calculus.)
10. Find all monic irreducible cubics in $\mathbb{Z}_2[x]$.
11. If f in $\mathbb{Z}_2[x]$ is irreducible and $\deg f \geq 1$, show that f has a nonzero constant term and has an odd number of terms. Is the converse true? Explain.
12. Let p be a monic quartic in $\mathbb{Z}_2[x]$. Show that p is irreducible in $\mathbb{Z}_2[x]$ if and only if (1) p has no root in \mathbb{Z}_2 and (2) $p \neq x^4 + x^2 + 1$. [Hint: Example 5.]
13. Show that a monic quintic p in $\mathbb{Z}_2[x]$ is irreducible if and only if (1) p has no root in \mathbb{Z}_2 , and (2) p is neither $x^5 + x^4 + 1$ nor $x^5 + x + 1$. [Hint: Exercise 10.]
14. Find all monic irreducible quadratics in $\mathbb{Z}_3[x]$.
15. Find a list of six quartics in $\mathbb{Z}_3[x]$ such that a monic quartic p is irreducible if and only if it has no root in \mathbb{Z}_3 and is not in the list. [Hint: Exercise 14.]
16. Show that there are $\frac{1}{2}p(p-1)$ monic irreducible quadratics in $\mathbb{Z}_p[x]$, where $p \in \mathbb{Z}$ is a prime. [Hint: There are p^2 monic quadratics; subtract the number that factor.]
17. If $p \in \mathbb{Z}$ is a prime, prove the converse of Example 7: If $p \not\equiv 3 \pmod{4}$, then $x^2 + 1$ is not irreducible over \mathbb{Z}_p . [Hint: Corollary to Theorem 8 §1.3.]
18. In each case factor f as a product of irreducible polynomials in $F[x]$.
 - (a) $f = 3x^4 + 2$, $F = \mathbb{Z}_5$
 - (b) $f = 3x^4 + 2$, $F = \mathbb{Z}_{11}$
 - (c) $f = x^3 + 2x^2 + 2x + 1$, $F = \mathbb{Z}_7$
 - (d) $f = x^3 + 2x^2 + 2x + 1$, $F = \mathbb{Z}_3$

- (e) $f = x^4 - x^2 + x - 1$, $F = \mathbb{Z}_{13}$
 (f) $f = x^4 - x^2 + x - 1$, $F = \mathbb{Z}_{17}$
19. Factor $x^5 + x^4 + 1$ as a product of irreducible polynomials in $\mathbb{Z}_2[x]$.
 20. Factor $x^5 + x^2 - x + 1$ as a product of irreducible polynomials in $\mathbb{Z}_3[x]$.
 21. Show that each polynomial is irreducible in $\mathbb{Q}[x]$.
- (a) $3x^3 + 5x^2 + x + 2$
 - (b) $5x^3 + 2x + 3$
 - (c) $x^3 + 9x^2 + x + 6$
 - (d) $x^3 + x^2 + 10x + 8$
22. Show that each polynomial is irreducible in $\mathbb{Q}[x]$.
- (a) $x^5 + 6x^4 + 12x + 15$
 - (b) $4x^5 + 28x^4 + 7x^3 - 28x^2 + 14$
23. In each case use the method of Example 13 to show that f is irreducible over \mathbb{Q} .
- (a) $f = x^4 + 2x - 1$
 - (b) $f = x^4 + 4x + 1$
 - (c) $f = x^4 + m$, where $m = 4k - 3$, k an integer
 - (d) $f = x^4 + 4mx + 1$, where m is an integer
24. Show that $f = x^4 + 4x^3 + 4x^2 + 4x + 5$ is irreducible over \mathbb{Q} by considering $f(x-1)$.
 25. If $p \in \mathbb{Z}$ is an odd prime, show that $f = 1 - x + x^2 - \dots + x^{p-1}$ is irreducible over \mathbb{Q} . [Hint: Example 13 with $x \mapsto x-1$.]
 26. Write $f_n = x^{n-1} + x^{n-2} + \dots + x + 1$.
- (a) Factor f_4 and f_6 into irreducible polynomials in $\mathbb{Q}[x]$.
 - (b) Show that f_n is not irreducible if $n \geq 2$ is not a prime (see Example 13).
27. If $p \in \mathbb{Z}$ is a prime and m is an integer, show that $x^p + p^2mx + (p-1)$ is irreducible over \mathbb{Q} . [Hint: Example 13.]
 28. Find a polynomial in $\mathbb{Z}[x]$ irreducible over \mathbb{Q} but not over \mathbb{Z}_2 , \mathbb{Z}_3 , \mathbb{Z}_5 , and \mathbb{Z}_7 .
 29. Show that $x^n - p$ is irreducible in $\mathbb{Q}[x]$ for all $n \geq 2$ and all primes $p \in \mathbb{Z}$. (Hence $\mathbb{Q}[x]$ has infinitely many irreducible polynomials of every degree ≥ 2 .)
 30. Show that $x^p - a$ is not irreducible in $\mathbb{Z}_p[x]$ for every $a \in \mathbb{Z}_p$.
 31. Let $F \subseteq K$ be fields and let f and g be polynomials in $F[x]$.
- (a) If f is irreducible in $K[x]$, show that it is irreducible in $F[x]$.
 - (b) If f and g are relatively prime in $F[x]$, show that they are relatively prime in $K[x]$. [Hint: Theorem 10(4).]
32. Is $x^4 + 1$ irreducible over \mathbb{R} ? Defend your answer. [Hint: Theorem 4.]
 33. If $p \in \mathbb{Z}$ is a prime, show that $x^2 + x + 1$ is irreducible over \mathbb{Z}_p if $p = 2, 5, 11, 17$, and not irreducible if $p = 3, 7, 13, 19$.
 34. Let $f = x^3 - 42x^2 + 35x + m$. Show that there are infinitely many integers m for which f is irreducible in $\mathbb{Q}[x]$. [Hint: Eisenstein.]
 35. In each case factor f into irreducible polynomials in $\mathbb{Q}[x]$.
- (a) $f = x^4 + 3x^3 + x^2 + 3x + 1$
 - (b) $f = x^4 + x^3 - 7x^2 + 3x - 2$
 - (c) $f = x^4 + 2x^3 - 2x^2 + 7x - 2$
 - (d) $f = x^4 - x^3 + 2x^2 - 3x + 2$
36. If m and p are integers with p prime, show that $x^4 + mx + p$ is irreducible in $\mathbb{Q}[x]$ if and only if it has no root in \mathbb{Q} .
 37. (a) Factor $x^5 + x + 1$ as a product of irreducible polynomials in $\mathbb{Q}[x]$.
 (b) Factor $x^5 + 3x + 1$ as a product of irreducible polynomials in $\mathbb{Q}[x]$.
 38. Let $F \subseteq K$ be fields and let f and $g \neq 0$ be in $F[x]$. If $f = qg$ for some $q \in K[x]$, show that actually $q \in F[x]$. [Hint: Division algorithm.]

39. In each case compute $d = \gcd(f, g)$, and express it in $F[x]$ as a linear combination of f and g .
- $f = x^2 + 2, g = x^3 + 4x^2 + x + 1; F = \mathbb{Z}_5$
 - $f = x^2 + 1, g = x^5 + x^4 + x^3 + x^2 + x + 1; F = \mathbb{Z}_2$
 - $f = x^2 - x - 2, g = x^5 - 4x^3 - 2x^2 + 7x - 6; F = \mathbb{Q}$
 - $f = x^3 + x - 2, g = x^5 - x^4 + 2x^2 - x - 1; F = \mathbb{Q}$
40. Prove the Corollary to Theorem 9.
41. Let f and g be monic in $F[x]$, F a field. Show that f divides g if and only if $\gcd(f, g) = f$. [Hint: Theorem 10.]
42. If F is a field, let $\gcd(f, g) = 1$ in $F[x]$. Mimic the proof in \mathbb{Z} to:
- Show that, if f and g both divide h , then fg divides h .
 - Show that, if f divides gh , then f divides h .
43. Let F be a field. A ring homomorphism $\sigma : F[x] \rightarrow F[x]$ is said to fix F if $\sigma(a) = a$ for all $a \in F$.
- If $b \in F$ and $\sigma(f) = f(x+b)$ then σ is a ring automorphism fixing F .
 - If $0 \neq a \in F$ and $\sigma(f) = f(ax)$ then σ is a ring automorphism fixing F .
 - If $\sigma : F[x] \rightarrow F[x]$ is any ring automorphism that fixes F , show that b and $a \neq 0$ exist in F such that $\sigma(f) = f(ax+b)$ for all f in $F[x]$.
44. Let F be a field, and let $t \mapsto \bar{t}$ is a ring automorphism $F \rightarrow F$. Given a polynomial $f = a_0 + a_1x + \cdots + a_nx^n$ in $F[x]$, define \bar{f} in $F[x]$ by $\bar{f} = \bar{a}_0 + \bar{a}_1x + \cdots + \bar{a}_nx^n$.
- Show that $f \mapsto \bar{f}$ is a ring automorphism $F[x] \rightarrow F[x]$.
 - If $\sigma : F[x] \rightarrow F[x]$ is any ring automorphism, show that there exist $a \neq 0$ and b in F , and an automorphism $t \mapsto \bar{t}$ of F , such that $\sigma(f) = \bar{f}(ax+b)$. [See Exercise 43.]

4.3 FACTOR RINGS OF POLYNOMIALS OVER A FIELD

If F is a field the similarity in Section 4.2 between the factorization theory in $F[x]$ and that in \mathbb{Z} continues. Every ideal of \mathbb{Z} has the form $n\mathbb{Z} = \langle n \rangle$, the factor ring $\mathbb{Z}/n\mathbb{Z}$ is easy to describe, and the generator n is uniquely determined if we insist that $n \geq 0$. This remains true in $F[x]$: Every ideal A of $F[x]$ is principal, that is $A = \langle h \rangle$ consists of all multiples of a polynomial $h \in F[x]$. Furthermore, h is uniquely determined by A if we ask that it is monic. Finally, we give a simple, explicit description of the factor ring $F[x]/\langle h \rangle$.

We begin with the fact that ideals in $F[x]$ are principal.

Theorem 1. *If F is a field every ideal A of $F[x]$ is principal. In fact, if $A \neq 0$, a uniquely determined monic polynomial h exists in $F[x]$ such that $A = \langle h \rangle$.*

Proof. If $A = 0$ then $A = \langle 0 \rangle$. If $A \neq 0$, it contains nonzero polynomials and hence contains monic polynomials (being an ideal). Among all the monic polynomials in A , choose h of minimal degree. Clearly, $\langle h \rangle \subseteq A$; we show that this is equality. If f is in A , the division algorithm (Theorem 4 §4.1) gives q and r in $F[x]$ such that $f = qh + r$ and either $r = 0$ or $\deg r < \deg h$. We show that $r = 0$ (so $f \in \langle h \rangle$).

Suppose $r \neq 0$, and let a be its leading coefficient. Then $a^{-1}r$ is monic and

$$a^{-1}r = a^{-1}[f - qh] = a^{-1}f - a^{-1}qh \in A.$$

But $\deg(a^{-1}r) = \deg r < \deg h$, contradicting the choice of h . So $r = 0$ and $A = \langle h \rangle$.

To prove uniqueness, suppose that $A = \langle k \rangle$, where k is also monic. Then $\langle k \rangle = \langle h \rangle$, so each of k and h divides the other. As both are monic, $k = h$ by Theorem 9 §4.2. \blacksquare

Hence both $F[x]$ and \mathbb{Z} are examples of **principal ideal domains**, that is integral domains in which every ideal is principal. We say more about these in Chapter 5.

If F is a field, Theorem 1 shows that the correspondence

$$h \leftrightarrow \langle h \rangle$$

is a bijection between the monic polynomials h in $F[x]$ and the nonzero ideals of $F[x]$. Note that $F[x] = \langle 1 \rangle$. Hence, our task in this section is to describe the factor rings $F[x]/\langle h \rangle$ in as much detail as possible, where h is any monic polynomial. Example 1 is an important special case, and the method of analysis serves as a prototype for the general case that follows.

Example 1. Describe the factor ring $R = \mathbb{R}[x]/A$, where $A = \langle x^2 + 1 \rangle$.

Solution. The elements of R are cosets $f + A$, $f \in \mathbb{R}[x]$, which we write as $\bar{f} = f + A$ for convenience. Hence, the operations in R are

$$\bar{f} + \bar{g} = \overline{f + g} \quad \text{and} \quad \bar{f} \cdot \bar{g} = \overline{fg}.$$

Given $f \in \mathbb{R}[x]$, the division algorithm gives q in $\mathbb{R}[x]$ such that

$$f = q(x^2 + 1) + (a + bx), \quad a, b \in \mathbb{R}.$$

Hence $f - (a + bx) \in A$, so $\bar{f} = \overline{a + bx} = \bar{a} + \bar{b}\bar{x}$. Thus R can be described as

$$R = \{\bar{a} + \bar{b}\bar{x} \mid a, b \in \mathbb{R}\}.$$

The ring axioms define the operations of R when the elements are presented in this way. The addition is easy

$$(\bar{a} + \bar{b}\bar{x}) + (\bar{c} + \bar{d}\bar{x}) = (\bar{a} + \bar{c}) + (\bar{b} + \bar{d})\bar{x}.$$

However, at first glance, the multiplication does not appear to be closed:

$$(\bar{a} + \bar{b}\bar{x})(\bar{c} + \bar{d}\bar{x}) = \overline{ac} + (\overline{ad} + \overline{bc})\bar{x} + \overline{bd}\bar{x}^2.$$

The problem is \bar{x}^2 . However, we have $x^2 + 1 \in A$, so $\bar{x}^2 + \bar{1} = \overline{x^2 + 1} = \bar{0}$. Thus, $\bar{x}^2 = -\bar{1}$ completes the description of the multiplication in R :

$$(\bar{a} + \bar{b}\bar{x})(\bar{c} + \bar{d}\bar{x}) = (\overline{ac} - \overline{bd}) + (\overline{ad} + \overline{bc})\bar{x}.$$

Does this look familiar? If we denote \bar{x} by a simpler symbol, say $\bar{x} = i$, then R looks like \mathbb{C} except for writing \bar{a} in place of a for all $a \in \mathbb{R}$. But even this difference is no problem: The map $a \mapsto \bar{a}$ is a one-to-one ring homomorphism $\mathbb{R} \rightarrow R$ (verify), so we may identify $\mathbb{R} \subseteq R$ as a subring by taking $\bar{a} = a$ for all $a \in R$. Finally then, our description of R takes the form

$$R = \{a + bi \mid a, b \in \mathbb{R}, i^2 = -1\}.$$

This is the ring \mathbb{C} of complex numbers, created before your eyes! (The only thing remaining to check is that $a + bi = c + di$ implies that $a = c$ and $b = d$, and this is left to the reader). \square

The analysis in Example 1 extends to the general case. For clarity we break the argument into a series of lemmas. The notation we use is as follows: Let F be a

field and let $h \in F[x]$ be a monic polynomial of degree $m \geq 1$. We write

$$A = \langle h \rangle = \{qh \mid q \in F[x]\}$$

for the principal ideal generated by h , and denote the (commutative) factor ring by

$$R = F[x]/A.$$

Then R consists of cosets $f + A$, $f \in F[x]$, and we write them as in Example 1:

$$\bar{f} = f + A.$$

In particular, $\bar{a} = a + A$ for each $a \in F$, and we adopt the symbol t for $x + A$:

$$t = \bar{x} = x + A.$$

The operations in R are

$$\bar{f} + \bar{g} = \overline{f + g} \quad \text{and} \quad \bar{f} \bar{g} = \overline{fg}.$$

Lemma 1. $R = \{\overline{a_0} + \overline{a_1}t + \cdots + \overline{a_{m-1}}t^{m-1} \mid a_i \in F\}$.

Proof. A typical element of R has the form \bar{f} , $f \in R[x]$. By the division algorithm, q exists in $F[x]$ such that

$$f = qh + (a_0 + a_1x + \cdots + a_{m-1}x^{m-1}), \quad a_i \in R.$$

Because $h \in A$, we have $\bar{h} = \bar{0}$ in R , so

$$\bar{f} = \overline{a_0 + a_1x + \cdots + a_{m-1}x^{m-1}} = \overline{a_0} + \overline{a_1}t + \cdots + \overline{a_{m-1}}t^{m-1}. \quad \blacksquare$$

Lemma 2. The map $\theta : F \rightarrow R$ given by $\theta(a) = \bar{a}$ is a one-to-one ring homomorphism.

Proof. The map θ is a homomorphism because $\overline{a+b} = \bar{a} + \bar{b}$, $\overline{ab} = \bar{a}\bar{b}$, and $\bar{1} = 1 + A$ is the unity of R . To see that θ is one-to-one, let $\theta(a) = \bar{0}$. Then $\bar{a} = \bar{0}$, so $a + A = 0 + A$, that is $a \in A$. If $a \neq 0$, then $A = F[x]$ because a is a unit in $F[x]$, so $1 \in A = \langle h \rangle$. This implies that $1 = hf$ for some f in $F[x]$ and hence that $\deg h = 0$, contrary to assumption. Thus, $a = 0$ and θ is one-to-one. \blacksquare

It follows from Lemma 2 that $\{\bar{a} \mid a \in F\} = \theta(F)$ is a subring of R that is isomorphic to F . Hence, we may identify $F = \theta(F) \subseteq R$ by taking $\bar{a} = a$ for all $a \in F$. This being done, Lemma 1 takes the form

$$R = \{a_0 + a_1t + \cdots + a_{m-1}t^{m-1} \mid a_i \in F\}.$$

Lemma 3 shows that the elements of R are uniquely represented in this way.

Lemma 3. If $a_0 + a_1t + \cdots + a_{m-1}t^{m-1} = b_0 + b_1t + \cdots + b_{m-1}t^{m-1}$ in R , then $a_i = b_i$ for each i .⁵⁹

Proof. The condition gives

$$(a_0 - b_0) + (a_1 - b_1)t + \cdots + (a_{m-1} - b_{m-1})t^{m-1} = 0.$$

Hence it suffices to show that $c_0 + c_1t + \cdots + c_{m-1}t^{m-1} = 0$, $c_i \in F$, implies that $c_i = 0$ for each i . To this end, write $k = c_0 + c_1x + \cdots + c_{m-1}x^{m-1}$. Then $k \in F[x]$ so it suffices to show that $k = 0$. First we have

$$\bar{k} = \overline{c_0} + \overline{c_1}\bar{x} + \cdots + \overline{c_{m-1}}\bar{x}^{m-1} = c_0 + c_1t + \cdots + c_{m-1}t^{m-1} = 0$$

⁵⁹Students of linear algebra will recognize this as showing that the set $\{1, t, t^2, \dots, t^{m-1}\}$ is linearly independent, and hence that it is a basis of R as an m -dimensional vector space over F .

in R . This means that $k \in A$, so $k = qh$ for some $q \in F[x]$. If $k \neq 0$, this gives

$$m - 1 \geq \deg k = \deg q + \deg h \geq \deg h = m,$$

a contradiction. Thus, $k = 0$ in $F[x]$, so $c_i = 0$ for all i , as required. ■

As in Example 1, the addition in R is straightforward in our new notation:

$$\begin{aligned} (a_0 + a_1t + \cdots + a_{m-1}t^{m-1}) + (b_0 + b_1t + \cdots + b_{m-1}t^{m-1}) \\ = (a_0 + b_0) + (a_1 + b_1)t + \cdots + (a_{m-1} + b_{m-1})t^{m-1}. \end{aligned}$$

However, the multiplication involves powers of t higher than $m - 1$. In the case of the complex numbers in Example 1, we wrote $t = i$, and the fact that $i^2 = -1$ enabled us to express the product in the form $a + bi$. In that situation, we had $h = x^2 + 1$, so the condition $i^2 = -1$ is $h(i) = 0$. This holds in general.

Lemma 4. *In the ring R , we have $h(t) = 0$.*

Proof. Write $h = c_0 + c_1x + \cdots + c_{m-1}x^{m-1} + x^m$. Recalling that $t = x + A$ and that we are writing $a = \bar{a} = a + A$ for all $a \in F$, we compute in R :

$$\begin{aligned} h(t) &= c_0 + c_1t + \cdots + c_{m-1}t^{m-1} + t^m \\ &= \overline{c_0} + \overline{c_1}\bar{x} + \cdots + \overline{c_{m-1}}\bar{x}^{m-1} + \bar{x}^m \\ &= \overline{c_0 + c_1x + \cdots + c_{m-1}x^{m-1} + x^m} \\ &= \bar{h} = \bar{0} = 0. \end{aligned}$$

The following theorem summarizes all the information we have gathered.

Theorem 2. *Let F be a field and let h be a monic polynomial in $F[x]$ of degree $m \geq 1$. Then the factor ring $F[x]/\langle h \rangle$ is given by*

$$\frac{F[x]}{\langle h \rangle} = \{a_0 + a_1t + \cdots + a_{m-1}t^{m-1} \mid a_i \in F; h(t) = 0\}.$$

Moreover, this representation of the elements of $F[x]/\langle h \rangle$ is unique:

$$a_0 + a_1t + \cdots + a_{m-1}t^{m-1} = b_0 + b_1t + \cdots + b_{m-1}t^{m-1}$$

holds if and only if $a_i = b_i$ for each i .⁶⁰

The formulation in Theorem 2 completely describes the ring $F[x]/\langle h \rangle$. Each element is uniquely represented in the form:

$$a_0 + a_1t + \cdots + a_{m-1}t^{m-1}, \quad a_i \in R \tag{*}$$

where $m = \deg h$. Addition and multiplication of such expressions are given by the ring axioms, the operations in F , and the requirement that h is monic and $h(t) = 0$. These conditions on h allow us to express t^m in terms of lower powers of t , and hence to reduce all products in R to the form (*), as guaranteed by Lemma 1. Thus, the multiplication depends in a crucial way on the polynomial h .

Example 1 describes the situation when $F = \mathbb{R}$ and $h = x^2 + 1$, and the ring $\mathbb{R}[x]/\langle x^2 + 1 \rangle$ turns out to be the ring of complex numbers \mathbb{C} . (In general it is the ring $F(i)$ mentioned in Section 3.2.) Examples 2–7 provide more illustrations.

⁶⁰The reader may have noticed that the discussion leading to Theorem 2 makes little use of the fact that F is a field. In fact, Theorem 2 is valid for any commutative ring in place of F (Exercise 30).

Example 2. If F is a field, describe the factor ring $R = F[x]/\langle x^2 \rangle$.

Solution. Theorem 2 applies with $h = x^2$ and $m = 2$. Hence

$$F[x]/\langle x^2 \rangle = \{a + bt \mid a, b \in F; t^2 = 0\}.$$

Thus, the addition in R is $(a+bt)+(c+dt) = (a+c)+(b+d)t$, as before. However, because $t^2 = 0$, the multiplication is $(a+bt)(c+dt) = ac + (ad+bc)t$.

For a specific instance, take $F = \mathbb{Z}_2 = \{0, 1\}$. Then $\mathbb{Z}_2[x]/\langle x^2 \rangle = \{0, 1, t, 1+t\}$ is a ring with four elements. Because $1+1=0$ in \mathbb{Z}_2 and $t^2=0$, the addition and multiplication tables are as follows:

$+$	0	1	t	$1+t$	\times	0	1	t	$1+t$
0	0	1	t	$1+t$	0	0	0	0	0
1	1	0	$1+t$	t	1	0	1	t	$1+t$
t	t	$1+t$	0	1	t	0	t	0	t
$1+t$	$1+t$	t	1	0	$1+t$	0	$1+t$	t	1

Example 3. Describe the ring $\mathbb{Z}_2[x]/\langle x^2 - 1 \rangle$.

Solution. This is Theorem 2 with $F = \mathbb{Z}_2$, $h = x^2 - 1$ and $m = 2$. Here $t^2 = 1$ so we use the notation $t = g$ because $G = \{1, g\}$ is then the cyclic group of order 2. Hence $\mathbb{Z}_2[x]/\langle x^2 - 1 \rangle = \{a + bg \mid a, b \in \mathbb{Z}_2; g^2 = 1\}$ and so consists of linear combinations of the elements of G with coefficients from \mathbb{Z}_2 . As such, it is called a **group ring**, and is usually denoted \mathbb{Z}_2G :

$$\mathbb{Z}_2G = \{a + bg \mid a, b \in \mathbb{Z}_2; g^2 = 1\}.$$

The addition is the same as in Example 2 (with $t = g$), but the multiplication is

$$(a + bg)(c + dg) = (ac + bd) + (ad + bc)g.$$

The addition and multiplication tables for \mathbb{Z}_2G are as follows:

$+$	0	1	g	$1+g$	\times	0	1	g	$1+g$
0	0	1	g	$1+g$	0	0	0	0	0
1	1	0	$1+g$	g	1	0	1	g	$1+g$
g	g	$1+g$	0	1	g	0	g	1	$1+g$
$1+g$	$1+g$	g	1	0	$1+g$	0	$1+g$	$1+g$	0

Note that the additive group of \mathbb{Z}_2G is isomorphic to the additive group of the ring $R = \mathbb{Z}_2[x]/\langle x^2 \rangle$ in Example 2, but the multiplications are different. Nonetheless, the map $\theta : \mathbb{Z}_2G \rightarrow R = \mathbb{Z}_2[x]/\langle x^2 \rangle$ given by $\theta(a+bg) = (a+b) + bt = a + b(1+t)$ is a ring isomorphism as the reader can verify. \square

Group rings FG can be constructed for any field F and group G , and are important in both the theory of rings and groups. In particular, if $g^n = 1$ and $G = \{1, g, g^2, \dots, g^{n-1}\}$ is the cyclic group of order n , then

$$FG = \{a_1 + a_1g + \dots + a_{n-1}g^{n-1} \mid a_i \in F \text{ and } g^n = 1\}$$

comes from Theorem 2 with $h = x^n - 1$.

Example 4. Consider the ring $R = \mathbb{Z}_2[x]/\langle x^3 + 1 \rangle$. Here $h = x^3 + 1$, so $m = 3$ and $t^3 + 1 = 0$ in Theorem 2. Hence

$$R = \{a + bt + ct^2 \mid a, b, c \in \mathbb{Z}_2; t^3 + 1 = 0\}.$$

Now $|R| = 8$ because (by Lemma 3) there are two independent choices for each of a, b , and c in forming $a + bt + ct^2$. Thus

$$R = \{0, 1, t, t^2, 1+t, 1+t^2, t+t^2, 1+t+t^2\}.$$

Because $\text{char } \mathbb{Z}_2 = 2$, we have $1+1=0$ and $t^3=1$ in R . A typical calculation is

$$(1+t)(1+t+t^2) = 1+2t+2t^2+t^3 = 1+0+0+1 = 0.$$

The reader should verify that both $1+t+t^2$ and $t+t^2$ are idempotents in R . \square

Example 5. Describe the ring $R = \mathbb{Q}[x]/\langle x^2 - 2 \rangle$.

Solution. Here $h = x^2 - 2$ and $m = 2$, so $R = \{a + bt \mid a, b \in \mathbb{Q}; t^2 = 2\}$ by Theorem 2. Clearly, this is (isomorphic to) the subring $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ of \mathbb{R} described in Example 4 §3.2. \square

In Example 4 §3.2, we showed directly that the ring $\mathbb{Q}(\sqrt{2})$ is a field. In the present context this property follows immediately from the next theorem and the fact that $x^2 - 2$ is irreducible over \mathbb{Q} .

Theorem 3. Let h be a monic polynomial of degree $m \geq 1$ in $F[x]$, where F is a field. The following conditions are equivalent:

- (1) $F[x]/\langle h \rangle$ is a field.
- (2) $F[x]/\langle h \rangle$ is an integral domain.
- (3) h is irreducible over F .

Proof. (1) \Rightarrow (2). This is clear; every field is an integral domain.

(2) \Rightarrow (3). For convenience write $A = \langle h \rangle$. If $h = fg$ is a factorization in $F[x]$, compute $(f+A)(g+A) = fg + A = h + A = 0 + A = 0$ in $F[x]/A$. By (2), either $f+A = 0$ or $g+A = 0$; that is $f \in A$ or $g \in A$. If $f \in A$, then $f = qh$ for some $q \in F[x]$. Hence $h = fg = qhg$, so (as $F[x]$ is an integral domain) $1 = qg$, which implies $\deg g = 0$. Similarly, $g \in A$ implies $\deg f = 0$. This proves (3).

(3) \Rightarrow (1). Let $f+A \neq 0$, where $f \in F[x]$. Then $f \notin A$, so h does not divide f . Let $d = \gcd(h, f)$. Then $d|h$ so, because h is irreducible and both d and h are monic, either $d = 1$ or $h = d$ (Corollary to Theorem 9 §4.2). But $h = d$ implies that $h|f$, contrary to $f+A \neq 0$. Hence $d = 1$, so (by Theorem 10 §4.2) u and v exist in $F[x]$ such that $1 = uh + vf$. Then $(v+A)(f+A) = 1 + A$ because $h \in A$, so $f+A$ is a unit in $F[x]/A$. This proves (1). \blacksquare

Note that Theorem 3 is the analogue for $F[x]$ of Theorem 7 §1.3 for \mathbb{Z} .

Example 6. If p is a prime, $p \equiv 3 \pmod{4}$, show that a field of p^2 elements exists.

Solution. It is a consequence of Fermat's theorem (see Example 7, §4.2) that $x^2 + 1$ is irreducible over \mathbb{Z}_p when $p \equiv 3 \pmod{4}$. Hence,

$$F = \mathbb{Z}_p[x]/\langle x^2 + 1 \rangle = \{a + bt \mid a, b \in \mathbb{Z}_p; t^2 = -1\}$$

is a field by Theorem 3. Moreover, in forming a typical element $a + bt$ of F we have p choices for a and then (by Lemma 3) p independent choices for b . Hence, there are p^2 choices in all, so $|F| = p^2$. The field F was denoted $\mathbb{Z}_p(i)$ in Section 3.2. \square

It is clear from the solution to Example 6 that, if h is monic and irreducible of degree m , then the field $F[x]/\langle h \rangle$ has exactly p^m elements. Moreover, it turns out

that a monic irreducible polynomial of degree m exists in $\mathbb{Z}_p[x]$ for every $m \geq 1$. Hence, we can construct a field of order p^m for all primes p and all integers $m \geq 1$. In fact, we obtain *every* finite field in this way. We return to this in Section 6.4.

We note in passing that $x^2 + 1$ fails to be irreducible over \mathbb{Z}_p if the prime p is not congruent to 3 modulo 4 (Corollary to Theorem 8 §1.3). In particular, $x^2 + 1$ is not irreducible over \mathbb{Z}_2 , and so will not yield a field (of order $4 = 2^2$) by our construction. This is not a major problem since $x^2 + x + 1$ is irreducible over \mathbb{Z}_2 .

Example 7. Construct a field of four elements.

Solution. The polynomial $x^2 + x + 1$ has no root in \mathbb{Z}_2 , and so is irreducible. Hence the required field is

$$F = \frac{\mathbb{Z}_2[x]}{\langle x^2 + x + 1 \rangle} = \{a + bt \mid a, b \in \mathbb{Z}_2; t^2 + t + 1 = 0\}.$$

Thus, $F = \{0, 1, t, 1+t\}$ and $t^2 = t + 1$ (as $1 + 1 = 0$ in \mathbb{Z}_2). The addition and multiplication tables are as follows.

+	0	1	t	$1+t$	\times	0	1	t	$1+t$
0	0	1	t	$1+t$	0	0	0	0	0
1	1	0	$1+t$	t	1	0	1	t	$1+t$
t	t	$1+t$	0	1	t	0	t	$1+t$	1
$1+t$	$1+t$	t	1	0	$1+t$	0	$1+t$	1	t

□

Let F be a field and let f be any polynomial of positive degree in $F[x]$. Then f has a monic irreducible factor $p \in F[x]$ by the unique factorization theorem (Theorem 12 §4.2), say $f = pg$. Given p , Theorem 3 shows that $E = F[x]/\langle p \rangle$ is a field that contains F as a subfield (after identifying each $a \in F$ with the coset $\bar{a} = a + \langle p \rangle$ in E). In addition, E contains an element t such that $p(t) = 0$ in E . Hence $f(t) = p(t)g(t) = 0$ in E , so t is a root of f in E . Calling a field E an **extension** of F when $F \subseteq E$, we can state this assertion compactly:

Theorem 4. Kronecker's Theorem. *If F is any field and f is any polynomial in $F[x]$ of positive degree, there is an extension field of F in which f has a root.*

Theorem 4 is fundamental to the algebraic study of fields. Note that it not only proves that the extension exists, but also gives a precise form for its elements. We treat this topic in detail in Chapter 6.

If $F = \mathbb{Q}$ in Theorem 4, then the fundamental theorem of algebra asserts that \mathbb{C} is an extension of \mathbb{Q} in which *any* polynomial of positive degree in $\mathbb{Q}[x]$ has a root. Hence, strictly speaking, we do not need Kronecker's Theorem in this case. But no purely algebraic proof of the fundamental theorem is known; that is, every proof involves a limiting process at some stage.

Exercises 4.3

Throughout these exercises F denotes a field.

- In each case find a monic polynomial h in $F[x]$ such that $A = \langle h \rangle$.
 - $A = \{f \in F[x] \mid \text{The constant coefficient of } f \text{ is zero}\}$
 - $A = \{f \in F[x] \mid \text{The sum of the coefficients of } f \text{ is zero}\}$

- (c) $A = \{f \in \mathbb{Z}_2[x] \mid f(0) = f(1) = 0\}$ [Hint: Theorem 10 §4.2.]
 (d) $A = \{f \in \mathbb{Z}_3[x] \mid f(0) = f(1) = f(2) = 0\}$ [Hint: Theorem 10 §4.2.]
2. In each case describe $R = F[x]/\langle h \rangle$ as in Theorem 2 and write out the addition and multiplication tables for R .
- (a) $h = x^2 + 1, F = \mathbb{Z}_2$
 - (b) $h = x^2 + x, F = \mathbb{Z}_2$
 - (c) $h = x^3 + 1, F = \mathbb{Z}_2$
 - (d) $h = x^2 - 1, F = \mathbb{Z}_3$
 - (e) $h = x^2, F = \mathbb{Z}_3$
 - (f) $h = x^2 - x + 1, F = \mathbb{Z}_3$
3. Construct a field of order 8 and write down the multiplication table.
4. Construct a field of order 9 and write down the multiplication table.
5. In each case construct a field of the given order.
- (a) 27
 - (b) 25
 - (c) 121
 - (d) 49
6. In each case determine all idempotents, nilpotents, and units in $R = F[x]/\langle h \rangle$.
- (a) $h = x^2 - x$
 - (b) $h = x^2$
7. In each case show that r is a unit in $R = F[x]/\langle h \rangle$ and exhibit the inverse. Use the notation of Theorem 2.
- (a) $r = 1 + t^2, F = \mathbb{Z}_{11}, h = x^3 + 1$
 - (b) $r = 1 + t - t^2, F = \mathbb{Z}_7, h = x^3 + x^2 - 1$
8. Because $x - a$ is irreducible over the field F , Theorem 3 asserts that $F[x]/\langle x - a \rangle$ is a field. Describe this field. How is it related to F ?
9. Find a subring of \mathbb{R} isomorphic to $\mathbb{Q}[x]/\langle x^3 - 2 \rangle$.
10. (a) Show that $\frac{F[x]}{\langle x^2 \rangle} \cong \left\{ \begin{bmatrix} a & b \\ 0 & a \end{bmatrix} \mid a, b \in F \right\}$, a subring of $M_2(F)$.
 (b) Show that $\frac{F[x]}{\langle x^3 \rangle} \cong \left\{ \begin{bmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{bmatrix} \mid a, b, c \in F \right\}$, a subring of $M_3(F)$.
 (c) Generalize.
11. Find a ring isomorphism $F[x]/\langle x^2 - x \rangle \rightarrow F \times F$.
12. Let $R = F[x]/\langle x^2 - 1 \rangle = \{a + bt \mid a, b \in F; t^2 = 1\}$. Show that $a + bt$ is a unit in R if and only if $a^2 \neq b^2$. [Hint: If $r = a + bt$ let $r^* = a - bt$, and $N(r) = rr^*$. Show that $(rs)^* = r^*s^*$, and hence that $N(rs) = N(r)N(s)$, for all $r, s \in R$.]
13. (a) Let $h = x^2 - vx - u$ in $F[x]$, where u and v are fixed in F . Define

$$S = \left\{ \begin{bmatrix} a & b \\ bu & a + bv \end{bmatrix} \mid a, b \in F \right\}.$$

 Show that S is a subring of $M_2(F)$ and that $F[x]/\langle h \rangle \cong S$.
 (b) Rework Exercises 10(a) and 11 in the light of (a).
 (c) If $h = x^2 + 1$ and $F = \mathbb{R}$, obtain a subring of $M_2(F)$ isomorphic to \mathbb{C} .
14. Let $E = F[x]/\langle p \rangle$, where p is irreducible over F . In each case factor p into linear factors in $E[x]$.
- (a) $p = x^3 + x + 1, F = \mathbb{Z}_2$
 - (b) $p = x^3 + x^2 + 1, F = \mathbb{Z}_2$
 - (c) $p = x^3 - x + 1, F = \mathbb{Z}_3$
 - (d) $p = x^3 - x^2 + 1, F = \mathbb{Z}_3$
15. If p is a monic irreducible quadratic in $F[x]$, show that p factors into linear factors over $E = F[x]/\langle p \rangle$.
16. (a) Assume that $2 \neq 0$ in F and that $m \in F$ is such that $x^3 - m$ is irreducible over F . Write $E = F[x]/\langle x^3 - m \rangle$. Show that $x^3 - m$ factors into linear factors in $E[x]$ if and only if -3 is a square in F . [Hint: A quadratic $x^2 + rx + s$ factors into linear factors over a field if and only if the discriminant $r^2 - 4s$ is a square in the field.]

- (b) Show that $x^3 - 2$ does not factor into linear factors over $E = \mathbb{Q}[x]/\langle x^3 - 2 \rangle$.
17. Let F be a finite field, say $F = \{a_1, a_2, \dots, a_n\}$. If $m = (x - a_1)(x - a_2) \cdots (x - a_n)$ and $A = \{f \in F[x] \mid f(a_i) = 0 \text{ for all } i = 1, 2, \dots, n\}$ denotes the set of all polynomials in $F[x]$ that annihilate F , show that $A = \langle m \rangle$.
18. Let A denote the set of all polynomials in $\mathbb{Z}[x]$ with even constant term. Show that A is an ideal of $\mathbb{Z}[x]$ that is not principal. (Hence, Theorem 1 fails for integral domains in general.)
19. If R is an integral domain for which every ideal of $R[x]$ is principal, show that R must be a field. [Hint: Exercise 18.]
20. Show that a field of order p^2 exists for every prime p . [Hint: Exercise 16 §4.2.]
21. (a) If $a^2 - 4b$ is not a square for a and b in a field F , show that $x^2 + ax + b$ is irreducible in $F[x]$.
(b) Show that the converse of (a) holds if $2 \neq 0$ in F .
22. Let f and g be nonzero polynomials in $F[x]$.
(a) Show that $A = \{uf + vg \mid u, v \in F[x]\}$ is an ideal of $F[x]$.
(b) Explain how Theorem 1 is related to Theorem 10 §4.2.
23. Polynomials f_1, f_2, \dots, f_m in $F[x]$ are called **relatively prime** if 1 is the only monic divisor of all of them in $F[x]$. Show that f_1, f_2, \dots, f_m are relatively prime if and only if $1 = q_1f_1 + q_2f_2 + \cdots + q_mf_m$ for some q_i in $F[x]$. [Hint: Theorem 1.]
24. Let f and g be two nonzero polynomials in $F[x]$. By Theorem 12 §4.2, monic irreducible polynomials p_1, p_2, \dots, p_r exist such that

$$\begin{aligned} f &= ap_1^{f_1}p_2^{f_2} \cdots p_r^{f_r}; \quad 0 \leq f_i \in \mathbb{Z}, a \in F \\ g &= bp_1^{g_1}p_2^{g_2} \cdots p_r^{g_r}; \quad 0 \leq g_i \in \mathbb{Z}, b \in F \end{aligned}$$

Here we take $f_i = 0$ if p_i does not occur in the factorization of f (and write $p_i^0 = 1$); with a similar convention for g . Define a polynomial

$$m = p_1^{\max(f_1, g_1)}p_2^{\max(f_2, g_2)} \cdots p_r^{\max(f_r, g_r)}.$$

Show that the following hold.

- (a) m is monic.
- (b) m is a common multiple of f and g .
- (c) If q is any common multiple of f and g , then m divides q .
- (d) m is uniquely determined by (a), (b), and (c).

Then m is called the **least common multiple** of f and g , denoted $m = \text{lcm}(f, g)$.

25. Given f and g in $F[x]$, let $d = \gcd(f, g)$ and $m = \text{lcm}(f, g)$ (Exercise 24). Show
(a) $\langle f \rangle + \langle g \rangle = \langle d \rangle$ (b) $\langle f \rangle \cap \langle g \rangle = \langle m \rangle$
26. (a) Let $A \neq F[x]$ be an ideal of $F[x]$, where F is a field. If $A \neq 0$, show that A is prime if and only if it is maximal.
(b) What happens if $A = 0$? Defend your answer.
27. Let F be a field and let h be a monic polynomial in $F[x]$. Show that $F[x]/\langle h \rangle$ has no nonzero nilpotent elements if and only if $h = p_1p_2 \cdots p_r$, where the p_i are distinct monic irreducible polynomials. [Hint: Theorem 12 §4.2.]
28. Let F be a field and let $h \neq 1$ be a monic polynomial in $F[x]$. Show that every element of $F[x]/\langle h \rangle$ is either a unit or a nilpotent if and only if $h = p^n$, where $n \geq 1$ and p is monic and irreducible.
29. Let F be a field and let $h = pq$ in $F[x]$, all polynomials being monic. If p and q are relatively prime in $F[x]$, show that $\frac{F[x]}{\langle h \rangle} \cong \frac{F[x]}{\langle p \rangle} \times \frac{F[x]}{\langle q \rangle}$. [Hint: Exercise 25 and Theorem 8 §3.4.]

30. Prove that Theorem 2 is valid as stated for any commutative ring R in place of the field F . Identify the places where the proofs of Lemmas 1, 2, 3, and 4 require modifications and make the required changes.
31. Let h be a monic polynomial of degree m in $F[x]$, F a field. Let $A = \langle h \rangle$, and write $R = F[x]/A$. Note that R can be written as $R = \{f(t) \mid f \in F[x]\}$, where $t = x + A$ as in Lemma 1.
- (a) If I is an ideal of R , show that there is a uniquely determined, monic divisor d of h in $F[x]$ such that

$$I = \{q(t)d(t) \mid q(t) \in R\} = \{f(t) \mid d \text{ divides } f \text{ in } F[x]\}.$$

Thus, $I = \langle d(t) \rangle$ is a principal ideal of R .

- (b) If $I_1 = \langle d_1(t) \rangle$, where d_1 is a monic divisor of h , show that $I \subseteq I_1$ if and only if d_1 divides d in $F[x]$.
- (c) If $h = db$, where b is (necessarily) monic, show that

$$I = \{f(t) \mid f(t)b(t) = 0\}.$$

This asserts that every ideal of R is an annihilator.

- (d) If $\deg d = m$ and $\deg h = n$, show that every element $f(t)$ of I is uniquely represented in the form:

$$f(t) = a_0d(t) + a_1td(t) + \cdots + a_{n-m-1}t^{n-m-1}d(t), \quad a_i \in F.$$

4.4 PARTIAL FRACTIONS

In calculus the first step in integrating a quotient $f(x)/g(x)$ of polynomial functions is to express it in a simpler form by expanding as a sum of **partial fractions**. Students learn to find such expressions in specific cases but the reason they exist in general usually remains a mystery. This is clarified in this section. We begin with Example 1, showing how to use the theorem.

Example 1. Expand $\frac{2x^2-x+1}{(x-1)^2(x^2+1)}$ as a sum of partial fractions.

Solution. The theorem that we are going to prove asserts that real numbers (called constants) a, b, c , and d exist such that

$$\frac{2x^2-x+1}{(x-1)^2(x^2+1)} = \frac{a}{x-1} + \frac{b}{(x-1)^2} + \frac{cx+d}{x^2+1}.$$

Once we know that they exist, we can routinely determine the constants. We multiply through by $(x-1)^2(x^2+1)$ to clear denominators:

$$2x^2 - x + 1 = a(x-1)(x^2+1) + b(x^2+1) + (cx+d)(x-1)^2. \quad (*)$$

We find the constant b quickly by evaluating at 1. The result is $2 = 2b$; $b = 1$. If we evaluate at 0, 2, and -1 , we get

$$\begin{aligned} 1 &= -a + b + d, \\ 7 &= 5a + 5b + 2c + d, \\ 4 &= -4a + 2b - 4c + 4d. \end{aligned}$$

The result is $a = d = \frac{1}{2}$, $b = 1$, $c = -\frac{1}{2}$. Note that we may also obtain equations in the constants by comparing coefficients of like powers of x on both sides of (*). For example, the coefficients of x^3 are $0 = a + c$. \square

We need Theorem 1 below in the proof of the main theorem, but it has independent interest (and is valid over an arbitrary ring).

Theorem 1. Let p be a monic polynomial in $R[x]$, R any ring. Given $f \in R[x]$, there exist uniquely determined polynomials r_0, r_1, \dots, r_m in $R[x]$ such that

$$f = r_0 + r_1 p + \cdots + r_m p^m$$

and, for each i , either $r_i = 0$ or $\deg r_i < \deg p$.

Proof. If $f = 0$ or $\deg f < \deg p$, then $f = r_0$ does it. Otherwise use induction on $\deg f$. By the division algorithm (Theorem 4 §4.1) write $f = qp + r_0$, where $r_0 = 0$ or $\deg r_0 < \deg p$. Then $q \neq 0$ so, as p is monic, $\deg f = \deg q + \deg p > \deg q$. Hence, by induction, $q = r_1 + r_2 p + \cdots + r_m p^{m-1}$ for some m , where $r_i = 0$ or $\deg r_i < \deg p$ for each i . The required representation of f follows. We leave the proof that it is unique as Exercise 1. \blacksquare

Example 2. Given $a \in R$, Theorem 1 asserts that each polynomial $f \in R[x]$ has an expansion of the form $f = a_0 + a_1(x - a) + a_2(x - a)^2 + \cdots + a_m(x - a)^m$, where $a_i \in R$ for each i .

Note that, if R is commutative and $2, 3, \dots$ are units in R , we can show that the coefficients a_i in Example 2 are given by $a_i = \frac{1}{i!} f^{(i)}(a)$, where $f^{(i)}$ is the i^{th} formal derivative of the function f (see Section 6.4). This result is called Taylor's theorem.

If F is a field, the field of quotients Q of the integral domain $F[x]$ consists of quotients $\frac{f}{g}$ of polynomials $f, g \in F[x]$, $g \neq 0$, called **rational forms** over F (see Theorem 5 §3.2). These forms are added and multiplied analogously to rational fractions. Working in Q , Theorem 1 enables us to prove the main result of this section.

Theorem 2. Partial Fraction Expansion. Let F be a field, and let $f, g \in F[x]$, where $g \neq 0$. Then the rational form $\frac{f}{g}$ has a unique expansion as a polynomial plus the sum of a number of rational forms $\frac{r}{p^k}$ where the following hold:

- (1) p is irreducible in $F[x]$.
- (2) p^k is a divisor of g and $k \geq 1$.
- (3) Either $r = 0$ or $\deg r < \deg p$.

Proof. We work in the field of quotients Q of $F[x]$. Given a rational form $\frac{f}{g} \in Q$, the division algorithm shows that $f = qg + f_1$, where $f_1 = 0$ or $\deg f_1 < \deg g$. Hence $\frac{f}{g} = q + \frac{f_1}{g}$ so, passing to f_1 , we may assume that $\deg f < \deg g$. Write $g = p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}$, where each p_i is irreducible in $F[x]$ and each $k_i \geq 1$. We need the following observation.

Claim. $\frac{f}{g} = \frac{h_1}{p_1^{k_1}} + \frac{h_2}{p_2^{k_2}} + \cdots + \frac{h_m}{p_m^{k_m}}$ for polynomials h_i in $F[x]$.

Proof. Use induction on m , the result being clear if $m = 1$. If $m > 1$ write $g_1 = p_2^{k_2} \cdots p_m^{k_m}$. Then g_1 and $p_1^{k_1}$ are relatively prime so (Theorem 10 §4.2) write $1 = ug_1 + sp_1^{k_1}$, where u and s are in $F[x]$. Then

$$\frac{f}{g} = \frac{f(ug_1 + sp_1^{k_1})}{p_1^{k_1} g_1} = \frac{fu}{p_1^{k_1}} + \frac{fs}{g_1}.$$

Now induction applies to $\frac{fs}{g_1}$, which proves the Claim.

Given the Claim, it remains to expand $\frac{h}{p^k}$ in the required form, where h and p are in $F[x]$, p is irreducible, and $k \geq 1$. To this end, use Theorem 1 to write

$$h = r_0 + r_1 p + r_2 p^2 + \cdots + r_m p^m$$

where, for each i , either $r_i = 0$ or $\deg r_i < \deg p$. Then $\frac{h}{p^k}$ has the desired form (possibly with a polynomial summand), proving the existence of the expansion. We omit the proof of uniqueness. ■

If it happens that $\deg f < \deg g$ in Theorem 2, the resulting expansion is the same except that there is no polynomial term. If p^k occurs in the factorization of the denominator g into irreducible factors, it yields terms

$$\frac{r_1}{p} + \frac{r_2}{p^2} + \cdots + \frac{r_k}{p^k}$$

in the partial fraction expansion. Because $r_i = 0$ or $\deg r_i < \deg p$ for each i , we have determined the form of this expansion, and only the coefficients of the r_i remain to be calculated. For example, if $\deg f < 7$, we have

$$\frac{f}{(x^2 + x + 1)^2(x + 1)^3} = \frac{ax + b}{x^2 + x + 1} + \frac{cx + d}{(x^2 + x + 1)^2} + \frac{r}{x + 1} + \frac{s}{(x + 1)^2} + \frac{t}{(x + 1)^3}$$

for an appropriate choice of the constants a, b, c, d, r, s , and t . We give one more example.

Example 3. Expand $\frac{x}{(x^2 + x + 1)^2(x + 1)}$ in partial fractions over \mathbb{R} .

Solution. Because $x^2 + x + 1$ and $x + 1$ are irreducible in $\mathbb{R}[x]$, the form of the expansion is

$$\frac{x}{(x^2 + x + 1)^2(x + 1)} = \frac{ax + b}{x^2 + x + 1} + \frac{cx + d}{(x^2 + x + 1)^2} + \frac{e}{x + 1}.$$

Clearing denominators gives

$$x = (ax + b)(x^2 + x + 1)(x + 1) + (cx + d)(x + 1) + e(x^2 + x + 1)^2.$$

Now Evaluating at -1 gives $e = -1$.

Comparing coefficients of x^4 gives $0 = a + e$.

Comparing coefficients of x gives $1 = a + 2b + c + d + 2e$.

Evaluating at 0 and yields $0 = b + d + e$.

Evaluating at 1 yields $1 = 6a + 6b + 2c + 2d + 9e$.

The solution is $a = c = d = 1$, $b = 0$, $e = -1$. □

The only irreducible polynomials in $\mathbb{R}[x]$ are linear and quadratic (Corollary to Theorem 4 §4.2), so Theorem 2 shows that every rational form over \mathbb{R} is a sum of terms of the types

$$\text{Polynomials, } \frac{ax + b}{(x^2 + rx + s)^k}, \quad \text{and} \quad \frac{a}{(x + r)^k}.$$

It turns out that all these forms (and hence every rational form) can be integrated by using only elementary functions.

Exercises 4.4

1. Prove the uniqueness in Theorem 1.
2. In each case, express the rational form as a sum of partial fractions over \mathbb{R} .
 - (a) $\frac{x^2 - x + 1}{x(x^2 + x + 1)}$
 - (b) $\frac{1}{(x-1)^2(x^2+2)}$
 - (c) $\frac{x+1}{x(x^2+1)^2}$
 - (d) $\frac{x^3+x+1}{(x^2+1)^2}$
3. Expand $\frac{1}{(x-u_1)(x-u_2)\cdots(x-u_n)}$ as a sum of partial fractions over F , where the u_i are distinct elements of the field F .
4. Using partial fractions, deduce that $\frac{n!}{x(x+1)\cdots(x+n)} = \sum_{k=0}^n \binom{n}{k} \frac{(-1)^k}{(x+k)}$.

4.5 SYMMETRIC POLYNOMIALS⁶¹

For any ring R we can iterate the process of forming a polynomial ring to construct the ring $R[x, y] = (R[x])[y]$. Every member of $R[x, y]$ has the form

$$f(x, y) = p_0 + p_1 y + p_2 y^2 + \cdots = \sum_{j \geq 0} p_j y^j,$$

where each p_j is in $R[x]$ and the sum is finite. If we write $p_j = \sum_{i \geq 0} a_{ij} x^i$, where the a_{ij} are in R , then $f(x, y)$ becomes a finite double sum:

$$\begin{aligned} f(x, y) &= \sum_{i \geq 0} \sum_{j \geq 0} a_{ij} x^i y^j \\ &= a_{00} + a_{10} x + a_{01} y + a_{20} x^2 + a_{11} xy + a_{02} y^2 + \cdots \end{aligned} \tag{*}$$

Moreover, each of x and y commutes with the other and with every element of R . Thus, we may interchange the role of x and y , that is

$$R[x, y] = R[y, x].$$

This is called the ring of polynomials in the two indeterminates x and y . The reader should verify that the representation in (*) is unique in the sense that

$$\sum a_{ij} x^i y^j = \sum b_{ij} x^i y^j, \quad \text{if and only if} \quad a_{ij} = b_{ij}, \quad \text{for all } i \text{ and } j.$$

If $R = \mathbb{R}$, the elements of $\mathbb{R}[x, y]$ are the familiar polynomial expressions from calculus and geometry.

If R is any ring, define the ring $R[x_1, \dots, x_n]$ of **polynomials** in the indeterminates x_1, \dots, x_n recursively as follows

$$R[x_1, \dots, x_n] = (R[x_1, \dots, x_{n-1}])[x_n], \quad \text{for } n \geq 2.$$

Hence, $R[x_1, x_2] = (R[x_1])[x_2]$ as above, $R[x_1, x_2, x_3] = (R[x_1, x_2])[x_3]$, and so on. With induction, Theorems 1 and 2 §4.1 give immediately

Theorem 1. *If R is any ring, then $R[x_1, \dots, x_n]$ is a ring. Moreover, if R is commutative or a domain, so also is $R[x_1, \dots, x_n]$.*

By induction, the indeterminates x_1, \dots, x_n commute with each other and with all elements of R . Hence, the order in which the x_i are adjoined to R is irrelevant; that is, for all permutations σ in the symmetric group S_n , we have

$$R[x_1, \dots, x_n] = R[x_{\sigma 1}, \dots, x_{\sigma n}].$$

⁶¹The results in this section are needed in Sections 6.6 and 10.3, and nowhere else.

Moreover, induction shows that each polynomial in $R[x_1, \dots, x_n]$ is a finite sum

$$f(x_1, \dots, x_n) = \sum a_{i_1 \dots i_n} x_1^{i_1} \cdots x_n^{i_n}, \quad a_{i_1 \dots i_n} \in R,$$

where the sum is taken over all n -tuples (i_1, \dots, i_n) with each $i_k \geq 0$, and where only finitely many coefficients $a_{i_1 \dots i_n}$ are nonzero. This representation is unique in the sense that

$$\sum a_{i_1 \dots i_n} x_1^{i_1} \cdots x_n^{i_n} = \sum b_{i_1 \dots i_n} x_1^{i_1} \cdots x_n^{i_n}$$

if and only if $a_{i_1 \dots i_n} = b_{i_1 \dots i_n}$ for all i_1, \dots, i_n . Again, this is easily established by induction.

In discussing a polynomial in $R[x_1, \dots, x_n]$, having names for the individual terms that make up the polynomial is useful. If $i_1 \geq 0, \dots, i_n \geq 0$ are integers, a polynomial of the form

$$m = m(x_1, \dots, x_n) = x_1^{i_1} \cdots x_n^{i_n}$$

is called a **monomial** in $R[x_1, \dots, x_n]$ and am , $a \in R$, is called a **monomial term**. If $a \neq 0$, the **degree** of am is defined to be $\deg(am) = i_1 + \cdots + i_n$. The **degree** of any nonzero polynomial in $R[x_1, \dots, x_n]$ is defined to be the maximum of the degrees of its nonzero monomial terms. A nonzero polynomial in $R[x_1, \dots, x_n]$ is called **homogeneous** if each of its monomial terms has the same degree. This notion of degree coincides with our earlier definition in $R[x]$. If $f \neq 0$ in $R[x_1, \dots, x_n]$ and k_1, k_2, \dots, k_m are the integers occurring as degrees of monomials in f , we can write f as $f = h_1 + h_2 + \cdots + h_m$, where each h_i is homogeneous of degree k_i (h_i is the sum of all monomial terms in f of degree k_i). These terms h_i are called the **homogeneous components** of f .

Example 1. In $\mathbb{R}[x, y, z]$, $\deg(2x^3yz^2) = 6$, $\deg(-xz) = 2$, and $\deg 1 = 0$. The polynomials $x + y$, $xy + y^2$, and $xyz + x^2y + xy^2 + z^3$ are homogeneous. However, $x^2 + 2yz + xz^2$ is not homogeneous but has two homogeneous components: $x^2 + 2yz$ and xz^2 .

Given $f = f(x_1, \dots, x_n)$ in $R[x_1, \dots, x_n]$ and a_1, \dots, a_n in the center $Z(R)$ of R , we evaluate $f(a_1, \dots, a_n)$ as in Section 4.1. In fact the evaluation map

$$R[x_1, \dots, x_n] \rightarrow R \quad \text{given by} \quad f(x_1, \dots, x_n) \mapsto f(a_1, \dots, a_n)$$

is a ring homomorphism. This result follows by induction on n : If $n = 1$, it is Theorem 5 §4.1; If $n > 1$, the map is a composite of the mappings

$$R[x_1, \dots, x_n] = (R[x_1, \dots, x_{n-1}])[x_n] \rightarrow R[x_1, \dots, x_{n-1}] \rightarrow R,$$

where the first map is evaluation at a_n and the second map is evaluation at (a_1, \dots, a_{n-1}) . Both are ring homomorphisms: the first by Theorem 5 §4.1, because a_n is central in $R[x_1, \dots, x_{n-1}]$, and the second by induction. Hence, the composite is a ring homomorphism.

If $n \geq 2$, a polynomial in $R[x_1, \dots, x_n]$ can have more than one monomial term of maximal degree. This means that the notion of degree is not as useful here as it is for polynomials in one indeterminate. What we need is a way of ordering the monomials themselves. If

$$p = x_1^{p_1} x_2^{p_2} \cdots x_n^{p_n} \quad \text{and} \quad q = x_1^{q_1} x_2^{q_2} \cdots x_n^{q_n}$$

are monomials, we write $p \leq q$ if and only if $p = q$ or $p_k < q_k$, where k is the smallest integer t with $p_t \neq q_t$. We write $p < q$ if $p \leq q$ but $p \neq q$, and in this case

we say that q is **higher** than p . This is a total ordering of the monomials (Exercise 19) called the **lexicographic** (or **dictionary**) **order**. Thus, the ordering of two monomials (written as $x_1^{i_1}x_2^{i_2}\cdots x_n^{i_n}$) is determined by the exponents in the first place from the left in which they differ, as for words in a dictionary.

Example 2. Order the set $\{x_1^2x_3^4x_4^2, x_1^2x_2^2x_3x_4^3, x_1^2x_3^2x_4, x_1^2x_2^2x_3\}$ of monomials in $R[x_1, x_2, x_3, x_4]$.

Solution. $x_1^2x_2^0x_3^2x_4 < x_1^2x_2^0x_3^4x_4^2 < x_1^2x_2^2x_3^1x_4 < x_1^2x_2^2x_3^1x_4^3$. \square

If $f \neq 0$ in $R[x_1, \dots, x_n]$, let p be the highest monomial appearing in f . If p has (nonzero) coefficient $a \in R$, then ap is called the **highest term** in f and is denoted $\text{ht}(f)$, and a is called the **highest coefficient** of f .

Example 3. If $f(x_1, x_2) = 4x_1x_2 - 3x_1x_2^2 + 3x_2^2$, then $\text{ht}(f) = -3x_1x_2^2$.

The next result is a generalization of Theorem 3 §4.1 and is needed in the proof of the Fundamental Theorem (Theorem 4 below). An element $a \in R$ is called a **nonzero divisor** if $ar = 0$ and $sa = 0$ can only happen in R if $r = 0$ and $s = 0$.

Theorem 2. Let f and g be nonzero polynomials in $R[x_1, \dots, x_n]$. If the highest coefficient of one of them is a non-zero-divisor, then $\text{ht}(fg) = \text{ht}(f) \cdot \text{ht}(g)$.

Proof. Let $\text{ht}(f) = rp$ and $\text{ht}(g) = sq$, where $r \neq 0 \neq s$ in R , and write $p = x_1^{p_1} \cdots x_n^{p_n}$ and $q = x_1^{q_1} \cdots x_n^{q_n}$. We must show that $\text{ht}(fg) = rspq$. With $rs \neq 0$ by hypothesis, it suffices to show that, if a and b are monomials in f and g , with either $a < p$ or $b < q$, then $ab < pq$. Write $a = x_1^{a_1} \cdots x_n^{a_n}$ and $b = x_1^{b_1} \cdots x_n^{b_n}$ and assume that $a < p$, say $a_k < p_k$, where k is minimal such that $a_k \neq b_k$. Because $b \leq q$, there are two cases.

- *Case 1.* $b = q$. Then $b_i = q_i$ for all i , so $a_k + b_k < p_k + q_k$, where k is minimal, showing that $ab < pq$.
- *Case 2.* $b < q$. Now let $b_l < q_l$, where l is minimal. If m is the smaller of l and k , the reader should verify that $a_m + b_m < p_m + q_m$, where m is minimal. Hence $ab < pq$ in this case, too. \blacksquare

Symmetric Polynomials

A polynomial $f(x_1, x_2, \dots, x_n)$ in $R[x_1, x_2, \dots, x_n]$ is called a **symmetric polynomial** if it is unchanged by any permutation of the indeterminates x_i :

$$f(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = f(x_1, x_2, \dots, x_n), \quad \text{for all } \sigma \text{ in } S_n.$$

Example 4. Every constant polynomial is symmetric.

Example 5. $\sum_{i \neq j} x_i x_j^2$ is symmetric. If $n = 3$, this is

$$x_1 x_2^2 + x_1 x_3^2 + x_2 x_3^2 + x_2 x_1^2 + x_3 x_1^2 + x_3 x_2^2.$$

Example 6. $p_k(x_1, \dots, x_n) = x_1^k + x_2^k + \cdots + x_n^k$ is symmetric for $k \geq 0$, and is called the **k -power symmetric polynomial**. Note that $p_0(x_1, \dots, x_n) = n$.

Example 7. $d(x_1, x_2, \dots, x_n) = \prod_{i < j} (x_i - x_j)^2$ is a symmetric polynomial, called the **discriminant** of the x_i . For example, if $n = 3$, this is

$$d(x_1, x_2, x_3) = (x_1 - x_2)^2 (x_1 - x_3)^2 (x_2 - x_3)^2.$$

Let $R[t, x_1, x_2, \dots, x_n]$ be a polynomial ring in $n+1$ indeterminates and consider the expressions:

$$\begin{aligned}(t - x_1)(t - x_2) &= t^2 - (x_1 + x_2)t + x_1 x_2, \\(t - x_1)(t - x_2)(t - x_3) &= t^3 - (x_1 + x_2 + x_3)t^2 \\&\quad + (x_1 x_2 + x_1 x_3 + x_2 x_3)t - x_1 x_2 x_3.\end{aligned}$$

If we regard these expressions as polynomials in t , the coefficients of powers of t are symmetric polynomials in the x_i , because permuting the x_i does not affect the left-hand side of the equations. The general definition is as follows.

The **elementary symmetric polynomials** $s_0, s_1, s_2, \dots, s_n$ in $R[x_1, \dots, x_n]$ are defined as follows:

$$\begin{aligned}s_k(x_1, x_2, \dots, x_n) &= \sum_{i_1 < i_2 < \dots < i_k} x_{i_1} x_{i_2} \cdots x_{i_k}, \quad \text{for any } k = 1, 2, \dots, n. \\s_0(x_1, x_2, \dots, x_n) &= 1\end{aligned}$$

Thus $s_k(x_1, x_2, \dots, x_n)$ is the sum of all distinct products of k of the indeterminates. For example,

$$\begin{aligned}s_1(x_1, x_2, \dots, x_n) &= x_1 + x_2 + \cdots + x_n, \\s_n(x_1, x_2, \dots, x_n) &= x_1 x_2 \cdots x_n.\end{aligned}$$

If $n = 4$, we have

$$\begin{aligned}s_2(x_1, x_2, x_3, x_4) &= x_1 x_2 + x_1 x_3 + x_1 x_4 + x_2 x_3 + x_2 x_4 + x_3 x_4, \\s_3(x_1, x_2, x_3, x_4) &= x_1 x_2 x_3 + x_1 x_2 x_4 + x_1 x_3 x_4 + x_2 x_3 x_4.\end{aligned}$$

Note that s_k is homogeneous of degree k for each $k = 1, 2, \dots, n$.

One of the main reasons for the importance of the elementary symmetric polynomials is the way they are related to the roots of a polynomial. For example,

$$(t - x_1)(t - x_2) = t^2 - (x_1 + x_2)t + x_1 x_2 = t^2 - s_1(x_1, x_2)t + s_2(x_1, x_2).$$

Since $s_0(x_1, x_2, \dots, x_n) = 1$, this expression generalizes as follows.

Theorem 3. Write $s_k = s_k(x_1, \dots, x_n)$ for $1 \leq k \leq n$. Then

$$(t - x_1)(t - x_2) \cdots (t - x_n) = t^n - s_1 t^{n-1} + s_2 t^{n-2} - \cdots \pm s_n = \sum_{k=0}^n (-1)^k s_k t^{n-k}.$$

Proof. The coefficient of t^n is $1 = s_0$. The expansion of the left-hand side is the sum of all products of n terms, one from each of the factors $t - x_i$. If $k \geq 1$, each product involving t^{n-k} has the form $t^{n-k}(-x_{i_1})(-x_{i_2}) \cdots (-x_{i_k})$, where $i_1 < i_2 < \cdots < i_k$. The sum of these terms is clearly $t^{n-k}(-1)^k s_k$. ■

It follows from the definition that the set S of all symmetric polynomials in $R[x_1, \dots, x_n]$ is a subring containing the constant polynomials. Hence every polynomial $f(s_1, \dots, s_n)$ in the elementary symmetric polynomials (with coefficients in R) is again in S . The fundamental theorem shows that *every* symmetric polynomial has this form.

Theorem 4. Fundamental Theorem of Symmetric Polynomials. Let R be any ring and let S denote the subring of all symmetric polynomials in $R[x_1, \dots, x_n]$. Then every member of S may be written in precisely one way as a polynomial $f(s_1, s_2, \dots, s_n)$ in the elementary symmetric polynomials $s_k = s_k(x_1, \dots, x_n)$, where $f(x_1, x_2, \dots, x_n)$ is in $R[x_1, \dots, x_n]$. Thus the map

$$f(x_1, x_2, \dots, x_n) \mapsto f(s_1, s_2, \dots, s_n)$$

is a ring isomorphism from $R[x_1, x_2, \dots, x_n]$ onto S .

Proof. Let $g = g(x_1, \dots, x_n) \neq 0$ be symmetric. If k_1, \dots, k_m are the (distinct) integers that occur as degrees of monomials in f , then $g = g_1 + \dots + g_m$, where g_i is homogeneous of degree k_i for each i . Given $\sigma \in S_n$ and a monomial $h(x_1, \dots, x_n)$, the fact that $h(x_1, \dots, x_n)$ and $h(x_{\sigma 1}, \dots, x_{\sigma n})$ have the same degree shows that each g_i is itself symmetric. Hence, we may assume that g is homogeneous.

Let g be symmetric and homogeneous with highest term $\text{ht}(g) = ap$, where $a \neq 0$ in R and $p = x_1^{m_1} \cdots x_n^{m_n}$. Consider the transposition $\sigma = (k \ k+1)$ in S_n , where $1 \leq k < n$. Because g is symmetric, it contains the monomial term aq , where $q = x_1^{m_1} \cdots x_k^{m_{k+1}} x_{k+1}^{m_k} \cdots x_n^{m_n}$. Hence $p > q$ by the choice of p , which means that $m_k \geq m_{k+1}$ for each k and hence that $m_1 \geq m_2 \geq \dots \geq m_n$. But, given nonnegative integers k_1, k_2, \dots, k_n , Theorem 2 implies that

$$\begin{aligned} \text{ht}[s_1^{k_1} s_2^{k_2} \cdots s_n^{k_n}] &= x_1^{k_1} (x_1 x_2)^{k_2} (x_1 x_2 x_3)^{k_3} \cdots (x_1 x_2 \cdots x_n)^{k_n} \\ &= x_1^{k_1+k_2+k_3+\cdots+k_n} x_2^{k_2+k_3+\cdots+k_n} \cdots x_{n-1}^{k_{n-1}+k_n} x_n^{k_n} \end{aligned}$$

Hence, the polynomial $g_1 = as_1^{m_1-m_2} s_2^{m_2-m_3} \cdots s_{n-1}^{m_{n-1}-m_n} s_n^{m_n}$ has the same highest term as g , and so $g - g_1$ either is 0 or has a lower highest term than g . Since it clearly suffices to show that $g - g_1$ is a polynomial in the s_i , we can repeat the process. A finite number of such repetitions yield g as a polynomial in the s_i .

Next, we prove the uniqueness of the representation. If in $R[x_1, x_2, \dots, x_n]$ some polynomial can be expressed in two ways as a polynomial in s_1, s_2, \dots, s_n , subtracting gives an equation

$$\sum_{k_i} a_{k_1 \cdots k_n} s_1^{k_1} \cdots s_n^{k_n} = 0, \quad (**)$$

where all coefficients are nonzero. Now the polynomial $s_1^{k_1} s_2^{k_2} \cdots s_n^{k_n}$ has highest monomial $x_1^{k_1+k_2+k_3+\cdots+k_n} x_2^{k_2+k_3+\cdots+k_n} \cdots x_n^{k_n}$, which uniquely determines the integers k_1, \dots, k_n . Consequently, distinct monomials $s_1^{k_1} s_2^{k_2} \cdots s_n^{k_n}$ in the s_i have distinct highest monomials in the x_i . Choose the highest x_i monomial arising in this way from the terms in (**). Then it occurs only once in (**) and with a nonzero coefficient. This contradicts the uniqueness of the representation of 0 in $R[x_1, \dots, x_n]$ as a linear combination of x_i monomials.

Finally, the mapping

$$R[x_1, \dots, x_n] \rightarrow R[x_1, \dots, x_n] \quad \text{given by} \quad f(x_1, \dots, x_n) \mapsto f(s_1, \dots, s_n)$$

has image S by the first part of this proof and is one-to-one by the uniqueness. Since the mapping is evaluation at s_1, \dots, s_n , it is a ring homomorphism because each s_i commutes with every element of R (all coefficients of s_i are 1). Hence, S is a subring isomorphic to $R[x_1, \dots, x_n]$. ■

The proof of Theorem 4 provides a method to actually express a symmetric homogeneous polynomial f as a polynomial in the s_i . If $ax_1^{m_1} \cdots x_n^{m_n}$ is the highest monomial term in f , subtract the term $as_1^{m_1-m_2} s_2^{m_2-m_3} \cdots s_{n-1}^{m_{n-1}-m_n} s_n^{m_n}$ from f , and repeat the procedure if the result is not a polynomial in the s_i . Example 8 demonstrates the method.

Example 8. Express $f(x_1, x_2) = x_1 x_2^3 + x_1^3 x_2$ in terms of elementary symmetric polynomials.

Solution. Here $n = 2$ and f is homogeneous with highest term $x_1^3x_2$. Hence,

$$f - s_1^{3-1}s_2^1 = (x_1x_2^3 + x_1^3x_2) - (x_1 + x_2)^2(x_1x_2) = -2x_1^2x_2^2 = -2s_2^2.$$

Hence, we are done with one iteration in this case, and $f = s_1^2s_2 - 2s_2^2$. \square

A method of *undetermined coefficients* is often easier to use than the technique in Example 8. We let f be symmetric and homogeneous of degree n in $R[x_1, x_2, \dots, x_n]$. The proof of Theorem 4 shows that f is a linear combination (with coefficients in R) of polynomials $s_1^{k_1}s_2^{k_2}\dots s_n^{k_n}$ with degree m , that is, with $k_1 + 2k_2 + \dots + nk_n = m$. If f is as in Example 8, then $m = 4$ and $n = 2$, so the s_i monomials in f have the form $s_1^{k_1}s_2^{k_2}$, where $k_1 + 2k_2 = 4$. Hence, f itself has the form:

$$f = as_1^4 + bs_1^2s_2 + cs_2^2, \quad a, b, \text{ and } c \text{ in } R.$$

Substituting $(x_1, x_2) = (1, 0)$ gives $a = 0$; then $(x_1, x_2) = (1, -1)$ gives $c = -2$; finally, $(x_1, x_2) = (1, 1)$ gives $b = 1$. Example 9 provides another illustration.

Example 9. Express $f(x_1, \dots, x_n) = \sum_{i \neq j} x_i^2x_j$ in terms of elementary symmetric polynomials.

Solution. Since f is homogeneous of degree 3, it has the form $f = as_1^3 + bs_1s_2 + cs_3$. Taking $(x_1, \dots, x_n) = (1, 0, \dots, 0)$ yields $a = 0$; then $(x_1, \dots, x_n) = (1, 1, 0, \dots, 0)$ gives $b = 1$; and, finally, $(x_1, \dots, x_n) = (1, 1, 1, 0, \dots, 0)$ gives $c = -3$. Hence $f = s_1s_2 - 3s_3$. \square

Note that the solution to Example 9 is based on the tacit assumption that $n \geq 3$ when expanding f (so s_3 can be written down). If $n = 2$, then $f(x_1, x_2) = x_1^2x_2 + x_1x_2^2 = (x_1 + x_2)(x_1x_2) = s_1s_2$, so the formula in Example 9 holds here too if $s_3(x_1, x_2) = 0$. But any valid formula in $R[x_1, x_2, x_3]$ reduces to a formula in $R[x_1, x_2]$ simply by taking $x_3 = 0$. Thus, $s_3(x_1, x_2, 0) = 0$, $s_2(x_1, x_2, 0) = s_2(x_1, x_2)$, and $s_1(x_1, x_2, 0) = s_1(x_1, x_2)$. Hence, the formula in Example 9 is valid even if $n = 2$.

The k -power polynomials $p_k(x_1, \dots, x_n) = x_1^k + x_2^k + \dots + x_n^k$ are symmetric and can be given in terms of s_1, \dots, s_n by formulas originating with Isaac Newton. The first three are

$$p_1 = s_1, \quad p_2 = s_1^2 - 2s_2, \quad \text{and} \quad p_3 = s_1^3 - 3s_1s_2 + 3s_3.$$

The first of these is clear and the others come from the following recursions.

Theorem 5. Newton's Identities. Let $p_k = p_k(x_1, \dots, x_n) = x_1^k + x_2^k + \dots + x_n^k$ denote the k -power symmetric polynomials. Then, for each $k > 1$,

$$p_k = p_{k-1}s_1 - p_{k-2}s_2 + \dots + (-1)^k p_1s_{k-1} + (-1)^{k+1}ks_k.$$

Note the coefficient k in the last term.

The proof is somewhat technical, and we present it at the end of this section.

Hence, given $p_1 = s_1$, the Newton identity with $k = 2$ gives

$$p_2 = p_1s_1 - 2s_2 = s_1^2 - 2s_2.$$

Then the case $k = 3$ gives $p_3 = p_2s_1 - p_1s_2 + 3s_3$, which yields

$$p_3 = s_1^3 - 3s_1s_2 + 3s_3.$$

Clearly, we can find p_4, p_5, \dots in the same way.

If σ is a permutation in S_n , the sign $\operatorname{sgn} \sigma$ of σ is defined by

$$\operatorname{sgn} \sigma = \begin{cases} 1, & \text{if } \sigma \text{ is even,} \\ -1, & \text{if } \sigma \text{ is odd.} \end{cases}$$

Then we can easily show (Exercise 29 §1.4) that $\operatorname{sgn} \sigma\tau = \operatorname{sgn} \sigma \cdot \operatorname{sgn} \tau$; that is, sgn is a group homomorphism $S_n \rightarrow \{1, -1\}$. The following class of polynomials is closely related to the symmetric polynomials.

A polynomial $f(x_1, \dots, x_n)$ in $R[x_1, \dots, x_n]$ is said to be **alternating** if

$$f(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = \operatorname{sgn} \sigma \cdot f(x_1, x_2, \dots, x_n) \quad \text{for all } \sigma \in S_n.$$

Examples include $(x_1 - x_2)x_1x_2$ and $(x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$. We characterize these alternating polynomials where R is a domain with characteristic not equal to 2.

As often happens, it is convenient to deal with a more general situation. Let $f(x_1, \dots, x_n)$ be a nonzero polynomial in $R[x_1, \dots, x_n]$, and suppose that a mapping $r : S_n \rightarrow R$ exists such that

$$f(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = r(\sigma) \cdot f(x_1, x_2, \dots, x_n) \quad \text{for all } \sigma \in S_n.$$

Thus, f is symmetric if $r(\sigma) = 1$ for all σ , and f is alternating if $r(\sigma) = \operatorname{sgn} \sigma$ for all σ . If R is a domain and $\varepsilon \in S_n$ is the identity permutation, it follows easily that

$$r(\varepsilon) = 1 \quad \text{and} \quad r(\sigma\tau) = r(\sigma) \cdot r(\tau)$$

for all σ and τ in S_n . In particular, if γ is a transposition ($\gamma^2 = \varepsilon$), then $r(\gamma) = \pm 1$. Since every permutation is a product of transpositions (Section 1.4), this shows that $r(\sigma) = \pm 1$ for all σ , and hence that

$$r : S_n \rightarrow \{1, -1\} \subseteq R$$

is a group homomorphism.

Let $K = \ker r = \{\sigma \in S_n \mid r(\sigma) = 1\}$. As $r(\sigma^2) = r(\sigma)^2 = 1$, we have $\sigma^2 \in K$ for all $\sigma \in S_n$. Hence, if $\sigma^3 = \varepsilon$, then $\sigma^{-1} = \sigma^2 \in K$, so $\sigma \in K$. In particular, K contains every 3-cycle and thus $A_n \subseteq K$. (A_n is generated by the 3-cycles by Lemma 2 §2.8.) But A_n has index 2 in S_n , so $A_n \subseteq K$ means that either $K = S_n$ or $K = A_n$. If $K = S_n$, then $r(\sigma) = 1$ for all σ and f is symmetric. If $K = A_n$, then $r(\sigma) = \operatorname{sgn} \sigma$ and f is alternating. This proves the first part of Theorem 6 below. To state it we need some terminology.

The polynomial $\Delta_n = \Delta_n(x_1, \dots, x_n) = \prod_{i < j} (x_i - x_j)$ is called the **alternator** of the variables x_i . Thus,

$$\begin{aligned} \Delta_2(x_1, x_2) &= (x_1 - x_2) \\ \Delta_3(x_1, x_2, x_3) &= (x_1 - x_2)(x_1 - x_3)(x_2 - x_3). \end{aligned}$$

These alternators clearly have the property that

$$\Delta_n(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = \pm \Delta_n(x_1, x_2, \dots, x_n), \quad \text{for all } \sigma \text{ in } S_n.$$

If R is a domain, the preceding discussion shows that Δ_n is either alternating or symmetric. But if $\sigma = (1 \ 2)$, then $\Delta_n(x_{\sigma 1}, \dots, x_{\sigma n}) = -\Delta_n(x_1, \dots, x_n)$ because σ negates $(x_1 - x_2)$ and permutes the other factors of Δ_n . Thus, Δ_n is not symmetric and so must be alternating.

Theorem 6. Let R be a domain, let $n \geq 2$, and let $0 \neq f \in R[x_1, \dots, x_n]$.

- (1) If for each $\sigma \in S_n$, $f(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = r(\sigma)f(x_1, x_2, \dots, x_n)$ for some $r(\sigma)$ in R , then f is either symmetric or alternating.
- (2) Assume that $\text{char } R \neq 2$. Then Δ_n is alternating, and f is alternating if and only if $f = \Delta_n g$ for some symmetric polynomial g .

Proof. It remains to prove (2). If $f = \Delta_n g$ with g symmetric, then f is alternating because Δ_n is alternating. Conversely, assume that f is alternating. If $\sigma = (1 \ 2)$ in S_n , then $f(x_2, x_1, x_3, \dots, x_n) = -f(x_1, x_2, x_3, \dots, x_n)$. Thus, $2f(x_1, x_1, x_3, \dots, x_n) = 0$, so $f(x_1, x_1, x_3, \dots, x_n) = 0$ (as R is a domain and $\text{char } R \neq 2$). Now view f as a polynomial in $S[x_1]$, where $S = R[x_2, \dots, x_n]$. Then x_2 is a root of f in S , so $f = (x_1 - x_2)h$ in $S[x_1]$ by the factor theorem (Theorem 6 §4.1). In the same way, x_3 is a root of f in S , so (as $x_3 \neq x_2$) it also is a root of h . This gives $f = (x_1 - x_2)(x_1 - x_3)k$ in $S[x_1]$, and eventually

$$f(x_1, \dots, x_n) = f = (x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_n)f_1(x_1, \dots, x_n). \quad (***)$$

We can now complete the proof by induction on $n \geq 2$. It is enough to show that $f = \Delta_n g$ because that implies that g is symmetric (both f and Δ_n are alternating). If $n = 2$, then (***)) reads $f = (x_1 - x_2)f_1 = \Delta_2 f_1$. In general, regard $f_1(x_1, \dots, x_n)$ in (***)) as in $T[x_2, \dots, x_n]$, where $T = R[x_1]$. Then f_1 is alternating because $(x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_n)$ is unchanged when x_2, x_3, \dots, x_n are permuted. By induction $f_1 = [\prod_{2 \leq i < j} (x_i - x_j)] g$, so $f = \Delta_n g$, as required. \blacksquare

We conclude with the promised proof of Newton's identities.

Proof of Theorem 5. Write

$$f(t) = (t - x_1)(t - x_2) \cdots (t - x_n) \text{ in } R[t, x_1, \dots, x_n].$$

Then Theorem 3 gives

$$f(t) = t^n - s_1 t^{n-1} + s_2 t^{n-2} + \cdots + (-1)^n s_n. \quad (1)$$

If $1 \leq i \leq n$, let $s_k^{(i)}$ denote the k th elementary symmetric function of the $n-1$ variables $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n$, where x_i is missing. Then we obtain

$$f(t) = (t - x_i)[t^{n-1} - s_1^{(i)} t^{n-2} + s_2^{(i)} t^{n-3} + \cdots + (-1)^{n-1} s_{n-1}^{(i)}].$$

Adding these equations for $i = 1, 2, \dots, n$ gives

$$\begin{aligned} \frac{f(t)}{t - x_1} + \frac{f(t)}{t - x_2} + \cdots + \frac{f(t)}{t - x_n} &= nt^{n-1} - \left[\sum_{i=1}^n s_1^{(i)} \right] t^{n-2} \\ &\quad + \left[\sum_{i=1}^n s_2^{(i)} \right] t^{n-3} - \cdots + (-1)^{n-1} \left[\sum_{i=1}^n s_{n-1}^{(i)} \right]. \end{aligned} \quad (2)$$

Now the product rule of differentiation shows that

$$(f_1 f_2 f_3 \cdots f_n)' = (f'_1 f_2 f_3 \cdots f_n) + (f_1 f'_2 f_3 \cdots f_n) + \cdots + (f_1 f_2 f'_3 \cdots f'_n).$$

Applying this rule to $f(t) = (t - x_1)(t - x_2) \cdots (t - x_n)$ shows that the left-hand side of (2) equals $f'(t)$. Then differentiating (1) term by term and comparing coefficients with (2) gives

$$(n - k)s_k = \sum_{i=1}^n s_k^{(i)}, \quad k = 1, 2, \dots, n - 1. \quad (3)$$

Now evaluate the sum on the right a different way: Group terms in the sum for s_k into those that involve x_i and those that do not, which gives $s_k = s_k^{(i)} + x_i s_{k-1}^{(i)}$. Hence

$$s_k^{(i)} = s_k - x_i s_{k-1}^{(i)}, \quad k = 1, 2, \dots, n-1.$$

Iterating gives $s_k^{(i)} = s_k - x_i s_{k-1} + x_i^2 s_{k-2}^{(i)}$. Continuing in this way, and using the fact that $s_0^{(i)} = 1$, yields

$$s_k^{(i)} = s_k - x_i s_{k-1} + x_i^2 s_{k-2} - \dots + (-1)^k x_i^k.$$

Sum this expression from $i = 1$ to n and use (3) to get

$$(n-k)s_k = ns_k - p_1 s_{k-1} + p_2 s_{k-2} - \dots + (-1)^k p_k,$$

which gives the k th Newton identity. ■

Exercises 4.5

1. Describe the units in $R[x_1, \dots, x_n]$.
2. In each case write f as the sum of its homogeneous components.
 - (a) $f(x, y, z) = x^3 + (x + yz)^2 + (x - y)(xz + z + 3)$
 - (b) $f(x, y, z) = (x - y)(x - z) + (x^2 + 1)(y^2 + xz) + 2(xz + 3)$
3. Exhibit a polynomial in $R[x, y]$ that is symmetric but not homogeneous, and one that is homogeneous but not symmetric.
4. If R is a domain and f and g are homogeneous of degrees m and n , show that fg is homogeneous of degree $m+n$.
5. Let $\theta : R \rightarrow S$ be a ring homomorphism and let $c_1, c_2, c_3, \dots, c_n$ be elements in the center of S . Show that there is a unique ring homomorphism $\bar{\theta} : R[x_1, \dots, x_n] \rightarrow S$ such that $\bar{\theta}(r) = \theta(r)$ for all $r \in R$ and $\bar{\theta}(x_i) = c_i$ for all i . We say that $\bar{\theta}$ is an extension of θ to $R[x_1, \dots, x_n]$.
6. Show that $f(x_1, \dots, x_n)$ is homogeneous of degree m in $R[x_1, \dots, x_n]$ if and only if $f(tx_1, \dots, tx_n) = t^m f(x_1, \dots, x_n)$ in $R[t, x_1, \dots, x_n]$, t another indeterminate.
7. In each case order the monomials lexicographically.
 - (a) $x_1 x_2^2 x_3, x_1 x_3, x_2^2 x_3, x_1^2 x_2$
 - (b) $x_2 x_3 x_4, x_1 x_3^2 x_4, x_2 x_3^2, x_1 x_4, x_2^2 x_4$
8. In each case, express the polynomial f in terms of elementary symmetric polynomials.
 - (a) $f(x_1, x_2) = \sum_{i \neq j} x_i^2 x_j^3$
 - (b) $f(x_1, x_2, x_3) = \sum_{i \neq j} x_i^2 x_j^3$
 - (c) $f(x_1, x_2, x_3) = \sum_{i \neq j \neq k \neq i} x_i^2 x_j^3 x_k$
 - (d) $f(x_1, x_2, x_3) = \sum_{i \neq j} x_i^4 x_j$
9. Show that the number of terms in $s_k(x_1, \dots, x_n)$ is $\binom{n}{k}$.
10. Show that the number of monomials of degree m in $R[x_1, \dots, x_n]$ is $\binom{m+n-1}{m}$. [Hint: How many ways can you place m zeros and $n-1$ ones in a row?]
11. Write p_4, p_5 , and p_6 in terms of elementary symmetric polynomials. What does the formula for p_5 say if $R = \mathbb{Z}_3$? Can you make a conjecture about p_q for any prime q ? If so, state it.
12. Using the Newton identities (or otherwise), express the following polynomials in x_1, x_2, \dots, x_n in terms of the elementary symmetric polynomials.

- (a) $f(x_1, \dots, x_n) = \sum_{i < j} (x_i - x_j)^2$ (b) $f(x_1, \dots, x_n) = \sum_{i < j} x_i^2 x_j^2$
 (c) $f(x_1, \dots, x_n) = \sum_{i < j} x_i^3 x_j^3$
13. Let the roots of $x^3 - 5x^2 + 4x - 3$ be u, v , and w .
 (a) Find the polynomial with roots u^2, v^2 , and w^2 .
 (b) Find the polynomial with roots $\frac{1}{u}, \frac{1}{v}$, and $\frac{1}{w}$.
14. Given $\sigma \in S_n$, define $\theta_\sigma : R[x_1, \dots, x_n] \rightarrow R[x_1, \dots, x_n]$ by

$$\theta_\sigma[f(x_1, \dots, x_n)] = f(x_{\sigma 1}, \dots, x_{\sigma n}).$$

 (a) Show that θ_σ is a ring automorphism of $R[x_1, \dots, x_n]$.
 (b) Show that $\sigma \mapsto \theta_\sigma$ is a group homomorphism $S_n \rightarrow \text{aut } R[x_1, \dots, x_n]$, which is one-to-one.
 (c) If $G \subseteq \text{aut } R[x_1, \dots, x_n]$ is a subgroup, show that $S_G = \{f \mid \theta(f) = f \text{ for all } \theta \in G\}$ is a subring of $R[x_1, \dots, x_n]$, called the ring of **G -symmetric polynomials**.
15. Let $f(x_1, \dots, x_n)$ be a polynomial in $\mathbb{Z}_p[x_1, \dots, x_n]$. If f has degree less than p in each indeterminate x_i , show that $f(a_1, \dots, a_n) \neq 0$ for some $a_i \in \mathbb{Z}_p$.
16. Find a symmetric polynomial $g(x, y)$ such that $x^m y^n - x^n y^m = \Delta_2 g(x, y)$. Assume that $m > n$.
17. Suppose that $p \in R[x]$ is odd; that is, $p(-x) = -p(x)$. If $f(x_1, \dots, x_n)$ is any alternating polynomial in $R[x_1, \dots, x_n]$, show that $f_1(x_1, \dots, x_n) = p[f(x_1, \dots, x_n)]$ is also alternating. If $f = \Delta_n g$, where g is symmetric, find a symmetric polynomial $g_1(x_1, \dots, x_n)$ such that $f_1 = \Delta_n g_1$.
18. Let S and A denote, respectively, the sets of symmetric and alternating polynomials in $R[x_1, \dots, x_n]$, and let $T = \{f + g \mid f \in S \text{ and } g \in A\}$. Assume that $n \geq 2$, R is a domain, and $\text{char } R \neq 2$. Show that T is a ring, Δ_n is central in T , $T = S + \Delta_n T$ as additive subgroups, $S \cap \Delta_n T = \Delta_n^2 S$, and $T/(\Delta_n T) \cong S/(\Delta_n^2 S)$ as rings.
19. Write n -tuples in \mathbb{N}^n as $a = (a_1, a_2, \dots, a_n)$. Define the **lexicographic order** or **dictionary order** on \mathbb{N}^n by $a \leq b$ if $a = b$ or $a_k < b_k$, where k is the smallest integer t with $a_t \neq b_t$.
 (a) Show that \leq is a **partial ordering** on \mathbb{N}^n ; that is $a \leq a$ for all a ; $a \leq b$ and $b \leq a$ imply that $a = b$; and $a \leq b$ and $b \leq c$ imply that $a \leq c$.
 (b) Show that \leq is a **total (or linear) ordering**; that is, $a \leq b$ or $b \leq a$ for all a and b in \mathbb{N}^n .
 (c) Show that \leq well **orders** \mathbb{N}^n ; that is, any nonempty set of n -tuples has a smallest element.
20. (a) Show that $G = \{a \in \mathbb{R} \mid -1 < a < 1\}$ is a group via $a * b = \frac{a+b}{1+ab}$.
 (b) Show that $x_1 * \dots * x_n = \frac{s_1+s_3+\dots+s_n}{1+s_2+\dots+s_{n-1}}$ if n is odd and $x_1 * \dots * x_n = \frac{s_1+s_3+\dots+s_{n-1}}{1+s_2+\dots+s_n}$ if n is even.

4.6 FORMAL CONSTRUCTION OF POLYNOMIALS

If R is any ring, we want to construct an indeterminate x over R , and so give precise meaning to the ring $R[x]$ of polynomials over R . We construct $R[x]$ as a subring of a larger ring S so that each expression $a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$ must be in S for any choice of $a_i \in R$. The elements a_i of R (and x) determine such an expression so, not surprisingly, we can construct S by using sequences from R .

A **sequence** from a ring R is a function $\alpha : \mathbb{N} \rightarrow R$. If we write $\alpha(m) = a_m$ for each $m \geq 0$, it is customary to display the sequence explicitly as a_0, a_1, a_2, \dots . We

will denote this sequence by

$$[a_m) = [a_0, a_1, a_2, \dots].$$

If $\beta : \mathbb{N} \rightarrow R$ is another sequence denoted $\beta(m) = b_m$, then $\alpha = \beta$ if and only if $a(m) = \beta(m)$ for all $m \geq 0$, that is, if and only if $a_m = b_m$ for all $n \geq 0$. In other words, two sequences equal when all the terms agree:

$$[a_m) = [b_m), \quad \text{if and only if} \quad a_m = b_m, \quad \text{for all } m \geq 0.^{62}$$

Now let S denote the set of *all* sequences from R :

$$S = \{[a_m) \mid a_m \in R \text{ for all } m \geq 0\}.$$

We are going to make S into a ring. We begin by defining addition on the set S :

$$[a_m) + [b_m) = [a_m + b_m).$$

It is an easy matter to verify that S is an abelian group with this addition. The zero element is the constant sequence $[0) = [0, 0, 0, \dots]$, and the negative of a sequence $[a_m)$ is $-[a_m) = [-a_m) = [-a_0, -a_1, -a_2, \dots]$.

The multiplication on S is **convolution**, defined as follows:

$$[a_m)[b_m) = [p_m) \quad \text{where} \quad p_m = \sum_{i+j=m} a_i b_j, \quad \text{for all } m \geq 0.$$

Hence $p_m = a_0 b_m + a_1 b_{m-1} + \dots + a_{m-1} b_1 + a_m b_0$ for each $m \in \mathbb{N}$. We leave to the reader the easy verification that the sequence $[1, 0, 0, \dots)$ is the unity for this multiplication. Next, we check associativity. Given three sequences $\bar{a} = [a_m)$, $\bar{b} = [b_m)$, and $\bar{c} = [c_m)$, we write $\bar{a}\bar{b} = [p_m)$, where $p_m = \sum_{i+j=m} a_i b_j$. Then $(\bar{a}\bar{b})\bar{c} = [p_m)[c_m) = [r_m)$, where

$$r_m = \sum_{t+k=m} p_t c_k = \sum_{t+k=m} (\sum_{i+j=t} a_i b_j) c_k = \sum_{i+j+k=m} (a_i b_j) c_k.$$

A similar calculation shows that $\bar{a}(\bar{b}\bar{c}) = [s_m)$, where $s_m = \sum_{i+j+k=m} a_i (b_j c_k)$. Hence, the associativity of the multiplication in S follows from that in R . A similar verification (which we also leave to the reader) shows that the distributive laws $\bar{a}(\bar{b} + \bar{c}) = \bar{a}\bar{b} + \bar{a}\bar{c}$, and $(\bar{b} + \bar{c})\bar{a} = \bar{b}\bar{a} + \bar{c}\bar{a}$ hold for all sequences \bar{a} , \bar{b} , and \bar{c} in S . Hence, S is a ring.

To construct $R[x]$ as a subring of S , we must first embed R as a subring of S . To this end, define $\theta : R \rightarrow S$ by $\theta(a) = [a, 0, 0, \dots)$ for all $a \in R$. One verifies that θ is a one-to-one ring homomorphism so R is isomorphic to the subring $\theta(R)$ of S . We identify these two copies of R by writing

$$a = \theta(a) = [a, 0, 0, 0, \dots),$$

which makes R into a subring of S . Finally, we define

$$x = [0, 1, 0, 0, \dots)$$

in S and observe that

$$ax = [a, 0, 0, \dots)[0, 1, 0, \dots) = [0, a, 0, \dots) = [0, 1, 0, \dots)[a, 0, 0, \dots) = xa$$

holds for all $a \in R$. Moreover, $ax^2 = [0, 0, a, 0, \dots)$, $ax^3 = [0, 0, 0, a, 0, \dots)$, \dots , so

$$a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n = [a_0, a_1, a_2, \dots, a_n, 0, 0, \dots)$$

for all $a_i \in R$. This result enables us to construct the polynomial ring $R[x]$.

⁶²This includes the construction of n -tuples as sequences $[a_m)$, where $a_m = 0$ for all $m \geq n$.

Theorem 1. Let R be any ring. There exists a ring S that contains R as a subring and contains an element x with the following properties:

- (1) $ax = xa$ for all $a \in R$.
- (2) If $a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = b_0 + b_1x + b_2x^2 + \cdots + b_nx^n$ in S , then $a_i = b_i$ for each $i \geq 0$.

Hence x is an indeterminant over R by (2), so the subring

$$R[x] = \{a_0 + a_1x + \cdots + a_nx^n \mid n \geq 0; a_i \in R \text{ for each } i\}$$

is the ring of polynomials over R and has all the properties required in Section 4.1. ■

The ring S is itself of interest. If x is as we defined it, we can write the sequences in S as

$$[a_m] = a_0 + a_1x + a_2x^2 + \cdots = \sum_{i=0}^{\infty} a_i x^i,$$

where *infinitely many* of the coefficients a_i may be nonzero. Thus, S is called the ring of **formal power series** over R , and is denoted $S = R[[x]]$. The polynomial ring $R[x]$ arises as the subring of all power series $\sum_{i=0}^{\infty} a_i x^i$ for which $a_i = 0$ for all but finitely many i .

Chapter 5

Factorization in Integral Domains

There still remain three studies suitable for free man. Arithmetic is one of them.

—Plato

We see therefore that ideal prime factors reveal the essence of complex numbers, make them transparent, as it were, and disclose their inner crystalline structure.

—Ernst Eduard Kummer

We have proved two unique factorization theorems: Every integer greater than one is uniquely a product of primes, and if F is a field every polynomial of positive degree is uniquely a product of an element of F times a product of monic irreducible polynomials. In this chapter, we characterize the integral domains for which a similar theorem holds (called unique factorization domains, or UFDs) and discuss some important classes of UFDs.⁶³

This theory has a long history and can be regarded as one of the original sources of modern abstract algebra. At the beginning of the nineteenth century, Gauss used the fact that the ring $\mathbb{Z}(i)$ (now called the gaussian integers) is a UFD to prove his law of biquadratic reciprocity, a method of determining when the congruence $x^4 \equiv b \pmod{n}$ has a solution. Inspired by the fact that i is a (fourth) root of unity, Kummer tried to extend Gauss' work by considering $\mathbb{Z}(w)$, where w is any complex root of unity. However, he discovered that $\mathbb{Z}(w)$ may not be a UFD. This observation had other implications. In 1847, Gabriel Lamé announced that he had solved one of the most famous problems in number theory, usually called Fermat's last theorem. It asserts that the equation $x^n + y^n = z^n$ has no solution in positive integers $x, y,$

⁶³Apart from Chapter 7, the material in this chapter is not essential elsewhere in this book.

and z for any integer $n \geq 3$. It is sufficient to prove this assertion if $n = p \geq 3$ is a prime. If w is a p th root of unity, Lamé had factored $x^p + y^p$ in $\mathbb{Z}(w)$ as

$$x^p + y^p = (x + y)(x + wy)(x + w^2y) \cdots (x + w^{p-1}y)$$

and then appealed to the (assumed) unique factorization in $\mathbb{Z}(w)$.

Kummer responded by proving that unique factorization *does* hold in $\mathbb{Z}(w)$ for what he called *ideal numbers*. This proof led to verification of Fermat's last theorem for many primes⁶⁴. However, Kummer's work had far greater significance for modern algebra because his ideal numbers were what we now call *ideals*. The idea was taken up by Dedekind, who characterized the integral domains in which every nonzero ideal is uniquely a product (suitably defined) of prime ideals.

5.1 IRREDUCIBLES AND UNIQUE FACTORIZATION

The higher arithmetic presents us with an inexhaustible storehouse of interesting truths... between which... we continually discover new and wholly unexpected points of contact.

—Carl Friedrich Gauss

Recall that a ring R is called an integral domain if it is commutative and $ab = 0$ in R implies that $a = 0$ or $b = 0$. In this section, we are concerned with factorization of elements in an integral domain R . We say that an element a of R is *factored* in R if it is equal to a product of two or more elements of R . Some factorizations are in a sense trivial. For example, $a = 1 \cdot a$ holds for all a . More generally, if u is a unit in R , then $a = u(u^{-1}a)$, and a factorization $a = ub$, where u is a unit, is called a **trivial factorization**. Such factorizations are of no interest, and we regard two factorizations $a = bc$ and $a = (ub)(u^{-1}c)$ as essentially the same.

As for \mathbb{Z} , if R is an integral domain and $a, b \in R$, we write $a|b$ if $b = ac$ for some $c \in R$. In this case, we say that a **divides** b or that a is a **divisor** of b . Verification of the following properties is easy.

- (1) $a|a$ for all $a \in R$.
- (2) If $a|b$ and $b|c$, then $a|c$.
- (3) If $a|b$ and $a|c$, then $a|(rb + sc)$ for all $r, s \in R$.

If m and n are nonzero integers, we can easily verify that both $m|n$ and $n|m$ hold if and only if $m = \pm n$, that is, if and only if $m = un$, where u is a unit of \mathbb{Z} . This holds in any integral domain R . Moreover, it is related to the set of principal ideals $\langle a \rangle = Ra$ generated by elements a of R .

Theorem 1. If R is an integral domain, the following are equivalent for $a, b \in R$:

- (1) $a|b$ and $b|a$.
- (2) $a = ub$ for some unit u in R .
- (3) $\langle a \rangle = \langle b \rangle$.

⁶⁴The “last theorem” remained open until 1997, when it was finally proved by Andrew Wiles.

Proof. (1) \Rightarrow (2). If $a|b$ and $b|a$, write $b = va$ and $a = ub$. If $a = 0$, then $b = va = 0$ too, so $a = 1b$. If $a \neq 0$, then $a = u(va) = (uv)a$ implies that $uv = 1$ because R is a domain. Thus, u is a unit.

(2) \Rightarrow (3). If $a = ub$, then $a \in Rb$, so $Ra \subseteq Rb$. Similarly, $b = u^{-1}a$ gives $Rb \subseteq Ra$. Hence, $Ra = Rb$, giving (3).

(3) \Rightarrow (1). If $\langle a \rangle = \langle b \rangle$, then $a \in \langle a \rangle = \langle b \rangle = Rb$. Hence, $b|a$; similarly, $a|b$. ■

Let R be an integral domain. If a and b are elements of R , we write

$$a \sim b \quad \text{if and only if} \quad a|b \text{ and } b|a.$$

In this case, a and b are said to be **associates** in R . Condition (3) in Theorem 1 implies immediately that the associate relation \sim is an equivalence on R :

- (1) $a \sim a$ for all $a \in R$.
- (2) If $a \sim b$, then $b \sim a$.
- (3) If $a \sim b$ and $b \sim c$, then $a \sim c$.

In this case, the equivalence class of $a \in R$ is

$$[a] = \{r \mid r \sim a\} = \{ua \mid u \text{ is a unit in } R\} = R^*a,$$

where, as usual, R^* denotes the group of units of R . In particular, $[0] = \{0\}$ and $[1] = R^*$. If $R = \mathbb{Z}$ and $n \in \mathbb{Z}$, then $[n] = \{n, -n\}$. If $R = F[x]$, where F is a field, and $f \in R$, then $[f] = \{af \mid 0 \neq a \in F\}$ because $F[x]^* = F^* = F \setminus \{0\}$.

Note that the associate relation \sim in an integral domain is compatible with divisibility and multiplication in the following sense:

- (1) If $a \sim a'$ and $b \sim b'$, then $a|b$ if and only if $a'|b'$.
- (2) If $a \sim a'$ and $b \sim b'$, then $ab \sim a'b'$.

These facts will be used frequently; we leave the verifications as Exercises 2 and 5.

Example 1. Show that $\sqrt{3} \sim (3 + 2\sqrt{3})$ in the integral domain

$$R = \mathbb{Z}(\sqrt{3}) = \{m + n\sqrt{3} \mid m, n \in \mathbb{Z}\}.$$

Solution. We have $3 + 2\sqrt{3} = (2 + \sqrt{3}) \cdot \sqrt{3}$, and $2 + \sqrt{3}$ is a unit in R (indeed, $(2 + \sqrt{3})(2 - \sqrt{3}) = 1$). □

We are interested in factorizations of elements of an integral domain R that are unique up to associates of the factors. Clearly, 0 must be excluded from consideration because $0 = 0 \cdot a$ holds for every $a \in R$. Also, if u is unit and $u = ab$, then both a and b are units; that is, all factorizations of a unit are trivial. Hence, we consider only nonzero nonunits. If such an element is factored nontrivially, one of the factors may have a nontrivial factorization. If this factorization is carried out, factors that can be further reduced may still remain. This process suggests consideration of those nonzero nonunits that, like the primes in \mathbb{Z} , admit no nontrivial factorization.

If R is an integral domain, $p \in R$ is called an **irreducible element**⁶⁵ (and is said to be **irreducible** in R) if it satisfies the following conditions:

- (1) $p \neq 0$ and p is not a unit.
- (2) If $p = ab$ in R , then a or b is a unit in R .

An element that is not irreducible is called **reducible**.

If $R = F[x]$, where F is a field, this definition agrees with the notion of an irreducible polynomial used in Section 4.2. However, the irreducibles in \mathbb{Z} are the elements of the form $\pm p$, where p is a prime. Note that a field has no irreducibles because no element is a nonzero nonunit.

Example 2. If $R = \mathbb{Z}(i) = \{m + ni \mid m, n \in \mathbb{Z}\}$ is the ring of gaussian integers, show that $p = 1 + i$ is irreducible in R .

Solution. Suppose that $p = ab$ in R . Taking absolute values gives $|a|^2|b|^2 = |p|^2 = 2$. Hence, $|a|^2 = 1$ or $|b|^2 = 1$ because $|a|^2$ and $|b|^2$ are positive integers. If $|a|^2 = 1$ and we write $a = m + ni$, where $m, n \in \mathbb{Z}$, then $m^2 + n^2 = |a|^2 = 1$. Hence, $a \in \{1, -1, i, -i\}$, so a is a unit in R . Similarly, $|b|^2 = 1$ implies that b is a unit. \square

Example 3. Let $R = \mathbb{Z}(\sqrt{-5}) = \{m + n\sqrt{-5} \mid m, n \in \mathbb{Z}\}$. Show that $p = 1 + \sqrt{-5}$ is irreducible in R .

Solution. If $a = m + n\sqrt{-5}$, we define the norm of a to be $N(a) = m^2 + 5n^2$. The reader can verify that $N(ab) = N(a)N(b)$ for a, b in R . Now suppose that $p = ab$ in R , so $6 = N(p) = N(a)N(b)$. Clearly, $N(a) = 2$ and $N(b) = 3$ are impossible, which means that $N(a) = 1$ or $N(b) = 1$. But then $a = \pm 1$ or $b = \pm 1$, so one of them is a unit. \square

The method in Examples 2 and 3 applies more generally, and we return to it in Section 5.2. First, we derive three useful conditions that make an element irreducible. The second is often taken as a definition of irreducibility.

Theorem 2. If R is an integral domain, the following conditions are equivalent for a nonzero nonunit p in R :

- (1) p is irreducible.
- (2) If $d|p$, then $d \sim 1$ or $d \sim p$.
- (3) If $p \sim ab$ in R , then $p \sim a$ or $p \sim b$.
- (4) If $p = ab$ in R , then $p \sim a$ or $p \sim b$.

Proof. (1) \Rightarrow (2). If $p = ad$, then d or a is a unit by (1), so $d \sim 1$ or $d \sim p$.

(2) \Rightarrow (3). If $p \sim ab$, then $b|p$, so $b \sim 1$ or $b \sim p$ by (2). In the first case, $p \sim a$.

(3) \Rightarrow (4). This is clear because $p = ab$ implies $p \sim ab$.

(4) \Rightarrow (1). If $p = ab$, then $p \sim a$ or $p \sim b$ by (4). If $p \sim a$, write $a = up$, where u is a unit. Then $p = ab = upb$, so $1 = ub$ (R is a domain). Thus, b is a unit. Similarly, $p \sim b$ implies that a is a unit, so (1) follows. \blacksquare

⁶⁵Irreducible elements are also called **atoms**.

An immediate consequence of Theorem 2 is that irreducibility is compatible with the associate relation \sim on R . More precisely,

If $p \sim q$, then p is irreducible if and only if q is irreducible.

We leave the proof as Exercise 16.

We can now give conditions on an integral domain R under which all nonzero nonunits can be factored in some way as a product of irreducibles.⁶⁶ For the moment, call a nonzero nonunit “bad” if it *cannot* be written as a product of irreducibles. Suppose a is “bad.” Then a is certainly not irreducible, so

$$a = x_1 a_1, \quad a \not\sim x_1 \quad \text{and} \quad a \not\sim a_1$$

by Theorem 2. Now at least one of x_1 and a_1 is “bad” (otherwise both are products of irreducibles, so a is not “bad”). Suppose that a_1 is “bad.” Then, as before,

$$a_1 = x_2 a_2, \quad a_1 \not\sim x_2 \quad \text{and} \quad a_1 \not\sim a_2,$$

where a_2 is “bad.” This process continues indefinitely. We have $a \in Ra_1 = \langle a_1 \rangle$, so $\langle a \rangle \subseteq \langle a_1 \rangle$. Similarly, $\langle a_1 \rangle \subseteq \langle a_2 \rangle$, $\langle a_2 \rangle \subseteq \langle a_3 \rangle, \dots$, and we obtain an ascending chain of principal ideals:

$$\langle a \rangle \subseteq \langle a_1 \rangle \subseteq \langle a_2 \rangle \subseteq \dots$$

Furthermore, $a \not\sim a_1, a_1 \not\sim a_2, \dots$, so (by Theorem 1) the chain is strictly increasing:

$$\langle a \rangle \subset \langle a_1 \rangle \subset \langle a_2 \rangle \subset \dots$$

Hence, any condition on R that rules out such strictly increasing chains guarantees that R contains no “bad” elements. Thus, the following definition is germane.

An integral domain R is said to satisfy the **ascending chain condition on principal ideals (ACCP)** if R contains no strictly increasing infinite ascending chain $\langle a_1 \rangle \subset \langle a_2 \rangle \subset \langle a_3 \rangle \subset \dots$ of principal ideals. The preceding argument proves

Theorem 3. *Let R be an integral domain that satisfies the ACCP. Then every nonzero nonunit in R is a product of irreducibles.*

The usefulness of Theorem 3 stems from the fact that the ACCP is easy to work with. One reason is the following useful alternative form of the condition.

Lemma 1. *The following conditions are equivalent for an integral domain R .*

- (1) R satisfies the ACCP.
- (2) For any ascending chain $\langle a_1 \rangle \subseteq \langle a_2 \rangle \subseteq \langle a_3 \rangle \subseteq \dots$ of principal ideals in R , an integer $n \geq 1$ exists such that $\langle a_n \rangle = \langle a_{n+1} \rangle = \dots$.

Proof. (1) \Rightarrow (2). Suppose $\langle a_1 \rangle \subseteq \langle a_2 \rangle \subseteq \dots$ but no n exists such that $\langle a_n \rangle = \langle a_{n+k} \rangle$ for all $k \geq 0$. Then $\langle a_1 \rangle \subset \langle a_{n_1} \rangle$ for some n_1 . Again, $\langle a_{n_1} \rangle \subset \langle a_{n_2} \rangle$ for some n_2 . By Theorem 3 §1.1, this continues to give $\langle a_1 \rangle \subset \langle a_{n_1} \rangle \subset \langle a_{n_2} \rangle \subset \dots$, contrary to (1). This proves (1) \Rightarrow (2). The proof that (2) \Rightarrow (1) is left as Exercise 20. \square

Example 4. Show that \mathbb{Z} satisfies the ACCP.

Solution. If $\langle a_1 \rangle \subseteq \langle a_2 \rangle \subseteq \langle a_3 \rangle \subseteq \dots$ in \mathbb{Z} , then $a_2|a_1, a_3|a_2, \dots$. Taking absolute values gives $|a_1| \geq |a_2| \geq |a_3| \geq \dots$. Since each $|a_i| \geq 0$ is an integer,

⁶⁶Meaning a product of one or more irreducibles.

$|a_n| = |a_{n+1}| = \dots$ must hold for some n . But then $a_{i+1} = \pm a_i$ for all $i \geq n$, so $\langle a_i \rangle = \langle a_{i+1} \rangle$ for all $i \geq n$, which is what we wanted. \square

More generally, we show in Section 5.2 that $\mathbb{Z}(w) = \{m + nw \mid m, n \in \mathbb{Z}\}$ satisfies the ACCP for any complex number w such that $w^2 \in \mathbb{Z}$ but $w \notin \mathbb{Q}$.

An argument similar to the proof of Lemma 1 (using degree in place of absolute value) shows that $F[x]$ satisfies the ACCP for any field F (Exercise 22). As F itself satisfies the ACCP (it has only two ideals!), this also follows from Theorem 4.

Theorem 4. *If R is an integral domain that satisfies the ACCP, then the ring $R[x]$ of polynomials over R also satisfies the ACCP.*

Proof. If not, let $\langle f_1 \rangle \subset \langle f_2 \rangle \subset \langle f_3 \rangle \subset \dots$ be a strictly increasing chain in $R[x]$. If a_i denotes the leading coefficient of f_i for each i , then $a_{i+1}|a_i$ in R because $f_{i+1}|f_i$, so $\langle a_1 \rangle \subseteq \langle a_2 \rangle \subseteq \langle a_3 \rangle \subseteq \dots$. By hypothesis, let $\langle a_n \rangle = \langle a_{n+1} \rangle = \dots$ for some $n \geq 1$; that is, $a_n \sim a_{n+1} \sim \dots$. If $m \geq n$, let $f_m = g f_{m+1}$, $g \in R[x]$. If b is the leading coefficient of g , then $a_m = b a_{m+1}$, so, as $a_m \sim a_{m+1}$, b is a unit in R . But g is not a unit in $R[x]$ because $\langle f_m \rangle \neq \langle f_{m+1} \rangle$, which means that $\deg g \geq 1$. Hence, $\deg f_m > \deg f_{m+1}$ by Theorem 3 §4.1. This is true for all $m \geq n$, so

$$\deg f_n > \deg f_{n+1} > \deg f_{n+2} > \dots$$

This is a contradiction since $\deg f_m$ is a nonnegative integer for each m . It follows that $R[x]$ satisfies the ACCP. \blacksquare

Hence, Example 4 shows that $\mathbb{Z}[x]$ satisfies the ACCP. More generally, if R is any integral domain satisfying the ACCP, then iterating Theorem 4 shows successively that the integral domains $R[x]$, $R[x, y] = (R[x])[y]$, $R[x, y, z] = (R[x, y])[z]$, \dots , all satisfy the ACCP.

Note that there exist integral domains in which the ACCP fails. For example, consider $R = \{n + xf \mid n \in \mathbb{Z}, f \in \mathbb{Q}[x]\}$, the set of all polynomials in $\mathbb{Q}[x]$ whose constant term is in \mathbb{Z} . Then R is an integral domain (subring of $\mathbb{Q}[x]$), but $\langle x \rangle \subset \langle \frac{1}{2}x \rangle \subset \langle \frac{1}{4}x \rangle \subset \dots$ as is readily verified. This interesting example is explored further in Exercise 40.

Unique Factorization Domains

Factorizations into irreducibles are much more useful when we know that they are unique up to associates of the factors. An integral domain R is called a **unique factorization domain (UFD)** if it satisfies the following conditions.

- (1) *Every nonzero nonunit in R is a product of irreducibles.*
- (2) *If $p_1 p_2 \cdots p_r \sim q_1 q_2 \cdots q_s$, where the p_i and the q_j are irreducibles, then $r = s$ and (after possible relabeling) $p_i \sim q_i$ for each i .*

Thus, Theorem 12 §4.2 shows that $F[x]$ is a UFD for any field F . Moreover, the field F is itself a UFD: Conditions (1) and (2) hold vacuously in this case because F contains no nonzero nonunits, and hence no irreducibles. Of course, \mathbb{Z} is the prototype example of a UFD. Note that Theorem 7 §1.2 proves unique factorization only for integers $n \geq 2$ in \mathbb{Z} . However, this theorem clearly extends to any integer $n \leq -2$ because $-p$ is irreducible for any prime p .

If $R = \mathbb{Z}$, scrutiny of the proof of Theorem 7 §1.2 shows that the uniqueness of the factorization into primes depends crucially on the primes $p \in \mathbb{Z}$ having the following property: If $p|ab$ in \mathbb{Z} , then $p|a$ or $p|b$ (Euclid's lemma). However, irreducibles in an arbitrary integral domain need not have this property, which leads us to yet another definition. An element p in an integral domain R is called a **prime element** of R (and is said to be **prime** in R) if it satisfies the following conditions:

- (1) $p \neq 0$ and p is not a unit.
- (2) If $p|ab$ in R , then $p|a$ or $p|b$.

Once again, being prime is compatible with the associate relation \sim , that is,

$$\text{If } p \sim q \text{ in } R \quad \text{then} \quad p \text{ is prime if and only if } q \text{ is prime.}$$

We leave the verification as Exercise 16.

Theorem 5. Every prime in an integral domain R is irreducible in R , but the converse fails for some integral domains R .

Proof. Let p be prime in R . If $p = ab$ in R , we must show that a or b is a unit. Clearly, $p|ab$, so $p|a$ or $p|b$ by hypothesis. If $p|a$, let $a = dp$. Then $a = d(ab)$, so, as $a \neq 0$ in the domain R , $1 = db$ and b is a unit. Similarly, $p|b$ implies that a is a unit, so p is irreducible. Example 5 shows that the converse can fail. ■

Example 5. If $R = \mathbb{Z}(\sqrt{-5}) = \{m + n\sqrt{-5} \mid m, n \in \mathbb{Z}\}$, show that $p = 1 + \sqrt{-5}$ is irreducible in R but not prime in R .

Solution. Example 3 shows that p is irreducible in R . To see that p is not a prime, observe that $2 \cdot 3 = 6 = (1 + \sqrt{-5})(1 - \sqrt{-5}) = p \cdot (1 - \sqrt{-5})$ in R . If p is a prime, this implies that $p|2$ or $p|3$ in R . Suppose that $p|2$, say $2 = qp$. Now write $N(m + n\sqrt{-5}) = m^2 + 5n^2$, much as in Example 3. Again N is multiplicative, that is, $N(xy) = N(x)N(y)$ holds for all x and y in R . Then $2 = qp$ gives

$$4 = N(2) = N(q)N(p) = N(q) \cdot 6,$$

which is impossible. Similarly, $p|3$ is impossible, so p is not a prime. □

The following analogue of Euclid's lemma (Theorem 6 §1.2) will be needed; the routine inductive proof extends and is left to the reader.

Lemma 2. Let p be a prime in an integral domain R . If $p|a_1a_2 \cdots a_n$ in R , then $p|a_i$ for some $i = 1, 2, \dots, n$.

Lemma 3. If R is a UFD, then every irreducible in R is prime.

Proof. Let $p \in R$ be irreducible. If $p|ab$ in R , write $ab = pd$. As R is a UFD, factor a, b , and d into irreducibles: $a = p_1 \cdots p_k$, $b = q_1 \cdots q_l$, and $d = r_1 \cdots r_m$. Then $pd = ab$ becomes $pr_1 \cdots r_m = p_1 \cdots p_k q_1 \cdots q_l$, so the uniqueness implies that either $p \sim p_i$ for some i or $p \sim q_j$ for some j . Hence, $p|a$ or $p|b$. ■

Hence, an element in a UFD is prime if and only if it is irreducible (using Theorem 5), so factorizations into irreducibles are actually factorizations into primes.

Many factorization properties of \mathbb{Z} extend automatically to any UFD; that is, the analogous proofs apply. We are going to discuss several of these properties and leave many details to the reader. We begin by describing divisors.

Let R be a UFD and let $a \in R$ be a nonzero nonunit. Then a can be written uniquely (up to associates) as a product

$$a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r},$$

where $a_i \geq 1$ for each i , each p_i is a prime in R by Lemma 3, and the p_i are nonassociates (that is, $p_i \not\sim p_j$ if $i \neq j$). Uniqueness means that the integers r, a_1, \dots, a_r are uniquely determined by a , as are the primes p_i up to associates. We claim that the divisors d of a are also determined uniquely (up to associates):

$$d|a \quad \text{if and only if} \quad d \sim p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r}, \text{ where } 0 \leq d_i \leq a_i \text{ for each } i. \quad (*)$$

Clearly, each d of this form is a divisor of a (possibly a unit).⁶⁷ Conversely, if $d|a$, then every prime divisor of d must be associated with one of the p_i by Lemma 2, so the prime factorization of d takes the form $d \sim p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r}$, $d_i \geq 0$. Similarly, if $a = db$ in R , then $b = p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r}$, $b_i \geq 0$, so

$$p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r} = a = db \sim p_1^{d_1+b_1} p_2^{d_2+b_2} \cdots p_r^{d_r+b_r}.$$

Uniqueness implies that $a_i = d_i + b_i$ for each i , so $d_i \leq a_i$, as asserted in (*).

Next, we define greatest common divisors and least common multiples in a UFD just as we did in \mathbb{Z} . Let s_1, s_2, \dots, s_n be elements of an integral domain R . An element d of R is called a **greatest common divisor** (**gcd**) of s_1, \dots, s_n , denoted $\gcd(s_1, \dots, s_n)$, if it satisfies the following conditions:

- (1) $d|s_i$ for each $i = 1, 2, \dots, n$.
- (2) If $r \in R$ and $r|s_i$ for each $i = 1, 2, \dots, n$, then $r|d$.

Analogously, $m \in R$ is called a **least common multiple** (**lcm**) of s_1, \dots, s_n , denoted by $\text{lcm}(s_1, \dots, s_n)$, if it satisfies

- (1) $s_i|m$ for each $i = 1, 2, \dots, n$.
- (2) If $r \in R$ and $s_i|r$ for each $i = 1, 2, \dots, n$, then $m|r$.

These definitions agree with those previously defined in \mathbb{Z} and $F[x]$, except that they are required to be positive in \mathbb{Z} and monic in $F[x]$. These extra conditions ensure uniqueness in \mathbb{Z} and $F[x]$, respectively, but no such device is available in an arbitrary UFD. However, $\gcd(s_1, \dots, s_n)$ and $\text{lcm}(s_1, \dots, s_n)$ are uniquely determined up to associates in any integral domain R :

$$\text{If } s_i \sim s'_i \text{ for each } i \quad \text{then} \quad \gcd(s_1, \dots, s_n) \sim \gcd(s'_1, \dots, s'_n).$$

A similar remark applies to least common multiples, and we leave the details to the reader (Exercise 25). Because we are ignoring the distinction between associates, we denote any greatest common divisor of s_1, \dots, s_n simply by $\gcd(s_1, \dots, s_n)$ and any least common multiple by $\text{lcm}(s_1, \dots, s_n)$.

Theorem 6 guarantees the existence of gcds and lcms of nonzero elements in any UFD (see Exercise 24).

⁶⁷Note that allowing zero exponents in (*) includes divisors where some prime p_i is missing.

Theorem 6. Let R be a UFD, and let a, b, c, \dots be a finite list of nonzero elements in R . If p_1, p_2, \dots, p_r are the nonassociated primes dividing one of a, b, c, \dots , write

$$a \sim p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}, \quad a_i \geq 0 \text{ in } \mathbb{Z},$$

$$b \sim p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r}, \quad b_i \geq 0 \text{ in } \mathbb{Z},$$

$$c \sim p_1^{c_1} p_2^{c_2} \cdots p_r^{c_r}, \quad c_i \geq 0 \text{ in } \mathbb{Z},$$

⋮

⋮

where an exponent is zero if the corresponding prime does not appear.⁶⁸ If we define $d_i = \min(a_i, b_i, c_i, \dots)$ and $m_i = \max(a_i, b_i, c_i, \dots)$ for each $i = 1, 2, \dots, r$, then

$$\gcd(a, b, c, \dots) \sim p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r} \quad \text{and} \quad \operatorname{lcm}(a, b, c, \dots) \sim p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r}.$$

Proof. The proof of Theorem 9 §1.2 carries over. ■

Warning. If R is a UFD and $d = \gcd(a, b)$ in R , it may *not* be possible to express d in the form $d = xa + yb$ for some x and y in R . This conclusion holds if $R = \mathbb{Z}$ or $R = F[x]$, F a field, but it need not hold in general. One condition guaranteeing that the condition holds is if every ideal of R is principal; we consider these “principal ideal domains” in Section 5.2.

Notwithstanding the warning, many properties of greatest common divisors familiar from the integers remain valid in any UFD, even though the method of proof is different. Example 6 illustrates this point. Compare the argument with the proof of Theorem 5(1) §1.2.

Example 6. If $a|c$, $b|c$, and $\gcd(a, b) = 1$ in a UFD, show that $ab|c$.

Solution. As in Theorem 6, choose primes p_1, \dots, p_r such that $a \sim p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$, $b \sim p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r}$, and $c \sim p_1^{c_1} p_2^{c_2} \cdots p_r^{c_r}$, where $a_i \geq 0$, $b_i \geq 0$, and $c_i \geq 0$ for all i . Then $a|c$ and $b|c$ give $a_i \leq c_i$ and $b_i \leq c_i$, respectively, for all i , whereas $\gcd(a, b) = 1$ means that $\min(a_i, b_i) = 0$ for all i . Thus, $a_i = 0$ or $b_i = 0$ for all i , so $a_i + b_i$ is a_i or b_i . In particular, $a_i + b_i \leq c_i$ holds for all i , whence $ab|c$. □

We need one more fact about gcds in an arbitrary integral domain.

Lemma 4. Let R be an integral domain and let a, b, c be nonzero elements of R . If $\gcd(a, b)$ and $\gcd(ca, cb)$ both exist in R , then $\gcd(ca, cb) \sim c \gcd(a, b)$.

Proof. Write $d = \gcd(a, b)$ and $d' = \gcd(ca, cb)$. Then $d|a$ and $d|b$, so $cd|ca$ and $cd|cb$. Hence, cd divides $\gcd(ca, cb) = d'$, say $d' = ucd$. So we show that u is a unit.

Write $ca = d'x$, $x \in R$, so $ca = ucdx$. Hence $a = udx$ as $c \neq 0$, so $ud|a$. Similarly, $ud|b$ so ud divides $\gcd(a, b) = d$. If $vud = d$, then $vu = 1$ as $d \neq 0$, so u is a unit. ■

We can now prove our characterization of unique factorization domains.

Theorem 7. The following are equivalent for an integral domain R :

- (1) R is a UFD.
- (2) R satisfies the ACCP, and $\gcd(a, b)$ exists for all nonzero $a, b \in R$.
- (3) R satisfies the ACCP, and every irreducible element in R is prime.

Proof. (1) \Rightarrow (2). Theorem 6 shows that gcds exist in R . Suppose, if possible, that $\langle a_1 \rangle \subset \langle a_2 \rangle \subset \langle a_3 \rangle \subset \cdots$ in R ; we look for a contradiction. We may assume that

⁶⁸If a is a unit, for example, then $a_i = 0$ for each i .

$a_1 \neq 0$. Moreover, a_1 is not a unit (because $\langle a_1 \rangle \neq R$). So let $a_1 = p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$, where p_i are nonassociated primes and $k_i \geq 1$ for each i . We have $a_i | a_1$ for all i , so $a_i \sim p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r}$ for $0 \leq d_j \leq k_j$. Thus, there are only finitely many nonassociated possibilities for the a_i , and so there must exist $m \neq n$ with $a_m \sim a_n$. But then $\langle a_m \rangle = \langle a_n \rangle$, a contradiction. Hence, R satisfies the ACCP.

(2) \Rightarrow (3). Let $p \in R$ be irreducible and assume that $p|ab$; we must show that $p|a$ or $p|b$. By (2), let $d = \gcd(a, p)$. Then $d|p$, so $d \sim p$ or $d \sim 1$ by hypothesis. In the first case, $p|a$ because $p \sim d$ and $d|a$. In the second case, $\gcd(a, p) \sim 1$. But then Lemma 4 gives $\gcd(ab, pb) \sim b$, so $p|b$ because $p|ab$ and $p|pb$. This proves (3).

(3) \Rightarrow (1). Given (3), each nonzero nonunit is a product of irreducibles by Theorem 3, so it remains to show that such factorizations are unique. If not, let

$$p_1 p_2 \cdots p_r \sim q_1 q_2 \cdots q_s$$

be distinct factorizations, where the p_i and q_i are irreducibles and $r + s$ is as small as possible. If $r = 1$, then $p_1 = q_1 q_2 \cdots q_s$ and it follows that $s = 1$ because p_1 is irreducible, a contradiction because the factorizations are distinct. So we may assume that $r \geq 2$ and $s \geq 2$. Then $p_1 | q_1 q_2 \cdots q_s$, so p_1 divides one of the q_j by Lemma 2. By relabeling the q_j , assume that $p_1 | q_1$. Since q_1 is irreducible, this implies that $p_1 \sim q_1$, whence $p_2 \cdots p_r \sim q_2 \cdots q_s$ are distinct factorizations, contradicting the minimality of $r + s$. ■

Unique Factorization in $R[x]$

We conclude this section with a proof that $R[x]$ is a UFD whenever R is a UFD. Theorem 7 guarantees that R satisfies the ACCP, and so the same is true of $R[x]$ by Theorem 4. Hence, by Theorem 7 again, all that remains is to show that irreducible polynomials in $R[x]$ are primes. This task is surprisingly difficult. Part of the problem is that irreducible elements of R remain irreducible as polynomials (of degree 0) in $R[x]$. The following definition helps to circumvent this difficulty.

If R is a UFD and f is a nonzero polynomial in $R[x]$, the greatest common divisor of the nonzero coefficients of f is called the **content** of f and is denoted $c(f)$. And f is called a **primitive polynomial** if $c(f) \sim 1$.

Example 7. In $\mathbb{Z}[x]$, $c(6 + 10x^2 + 15x^3) = 1$ but $c(6 + 9x^2 + 15x^3) = 3$.

If $c = c(f)$ is the content of a nonzero polynomial f , then c divides every coefficient of f and so $f = cf_1$, where f_1 is a polynomial uniquely determined up to associates by f . Moreover f_1 is primitive, which is the first part of Lemma 5. We leave the proof as Exercise 33.

Lemma 5. Let R be a UFD and let $f \neq 0$ be a polynomial in $R[x]$.

- (1) f can be written as $f = c(f) f_1$, where $f_1 \in R[x]$ is primitive.
- (2) If $0 \neq a \in R$, then $c(af) \sim a c(f)$.

Lemma 6. If R is a UFD and $p \in R[x]$ is irreducible with $\deg p \geq 1$, then p is primitive.

Proof. Write $p = cp_1$, where $c = c(p)$. Then c or p_1 is unit in $R[x]$ because p is irreducible. But p_1 is not a unit because $\deg p_1 \geq 1$, so c is a unit. Thus, $c \sim 1$, which means that p is primitive. □

The following theorem, first proved by Gauss at the end of the eighteenth century, is the key to proving that $R[x]$ is a UFD whenever R has this property.

Theorem 8. Gauss' Lemma. *Let R be a UFD. If $f \neq 0$ and $g \neq 0$ in $R[x]$, then*

$$c(fg) \sim c(f)c(g).$$

In particular, the product of primitive polynomials is primitive.

Proof. Let $f = c(f) \cdot f_1$ and $g = c(g) \cdot g_1$, where f_1 and g_1 are primitive. Then

$$c(fg) \sim c[c(f)c(g)f_1g_1] \sim c(f)c(g)c(f_1g_1)$$

by Lemma 5, so it suffices to prove the result when f and g are primitive. Hence, assume that $c(f) \sim 1 \sim c(g)$ and suppose that fg is not primitive. Then some prime p divides each coefficient of fg . Write $f = a_0 + a_1x + \dots$ and $g = b_0 + b_1x + \dots$. Because f and g are primitive, p does not divide every a_i (or every b_j), so $n \geq 0$ and $m \geq 0$ exist such that

- p does not divide a_n , but $p|a_i$ for $0 \leq i < n$, and
- p does not divide b_m , but $p|b_j$ for $0 \leq j < m$.

The coefficient of x^{m+n} in fg is $c = \sum_{i+j=m+n} a_i b_j$. Thus, $p|c$ and p divides every term $a_i b_j$ except possibly $a_n b_m$. But then $p|a_n b_m$ too so $p|a_n$ or $p|b_m$ because p is prime. This contradiction proves Gauss' lemma. \blacksquare

Our first use of Gauss' lemma is to prove Theorem 9, which, although useful in itself, is needed in the proof of our main result (Theorem 10).

Theorem 9. *Let R be a UFD with field of quotients F . Regard $R \subseteq F$ as a subring of F as usual. If $p \in R[x]$ is irreducible in $R[x]$, then p is irreducible in $F[x]$.*

Proof. Let p be irreducible in $R[x]$ and assume that $p = gh$ in $F[x]$. If a and b are the products of the denominators of the coefficients of g and h , then $g_1 = ag$ and $h_1 = bh$ are in $R[x]$, and $abp = g_1h_1$ is a factorization in $R[x]$. Moreover, p is primitive in $R[x]$ by Lemma 6, so Gauss' lemma gives

$$ab \sim ab c(p) \sim c(abp) = c(g_1h_1) \sim c(g_1)c(h_1). \quad (**)$$

Now write $g_1 = c(g_1)g_2$ and $h_1 = c(h_1)h_2$, where g_2 and h_2 are primitive in $R[x]$. Hence, $abp = g_1h_1 = c(h_1)c(g_1)h_2g_2$, so $(**)$ implies that $p \sim h_2g_2$ in $R[x]$. But then h_2 or g_2 is a unit in $R[x]$. If $g_2 = u$ is a unit in R , then $bg = g_1 = c(g_1)g_2 = c(g_1)u$, so $g = b^{-1}c(g_1)u$ is a unit in $F[x]$. Similarly, if $h_2 \in R^*$, then $h \in F[x]^*$. \blacksquare

Note that the converse of Theorem 9 is not true. For example, $3(x^2 + 1)$ is irreducible in $\mathbb{Q}[x]$ but not in $\mathbb{Z}[x]$.

We can now prove the most important theorem of this section.

Theorem 10. *If R is a UFD, the polynomial ring $R[x]$ is also a UFD.*

Proof. By Theorems 4 and 7, it suffices to show that every irreducible p in $R[x]$ is prime. Accordingly, assume that $p|fg$ in $R[x]$; we must prove that $p|f$ or $p|g$.

Claim 1. It suffices to prove that $p|f$ or $p|g$ when f and g are primitive. \blacksquare

Proof. Let $hp = fg$, where $h \in R[x]$. By Lemma 5, write $f = af_1$, $g = bg_1$, and $h = dh_1$, where a , b , and d are in R and f_1 , g_1 , and h_1 are primitive in $R[x]$. Because p is also primitive (by Lemma 6), Gauss' lemma gives

$$d \sim c(h) \sim c(hp) \sim c(fg) \sim c(f)c(g) \sim ab.$$

Hence, $h_1p \sim f_1g_1$ because $dh_1p = hp = fg = abf_1g_1$. Hence, $p|f_1g_1$ so, as f_1 and g_1 are primitive, our assumption implies that $p|f_1$ or $p|g_1$. If $f_1 = kp$, then $(ak)p = af_1 = f$, so $p|f$. Similarly, $p|g_1$ implies $p|g$, proving the Claim 1.

So assume that $hp = fg$, where f and g are primitive in $R[x]$. Let F denote the field of quotients of R and, as usual, regard $R \subseteq F$ as a subring of F . Then $p|fg$ in $F[x]$, so, as p is irreducible in $F[x]$ by Theorem 9, Theorem 11 §4.2 gives $p|f$ or $p|g$ in $F[x]$, say $f = kp$, $k \in F[x]$. If d is the product of all denominators of nonzero coefficients of k , then $g_0 = dk \in R[x]$ and we have $df = g_0p$. But f is now assumed to be primitive, so Gauss' lemma gives

$$d \sim c(df) \sim c(g_0p) \sim c(g_0)c(p) \sim c(g_0).$$

If we write $g_0 = c(g_0)g_1$ where $g_1 \in R[x]$, then $df = g_0p = c(g_0)g_1p$. As $d \sim c(g_0)$, it follows that $f \sim g_1p$, so $p|f$ in $R[x]$, as required. ■

In particular, $\mathbb{Z}[x]$ is a UFD, a result first proved by Gauss.

If R is a UFD, then $R[x]$ is also a UFD by Theorem 10, so the theorem shows that the ring $R[x, y] = (R[x])[y]$ of polynomials in two commuting indeterminates is also a UFD. More generally, define the ring $R[x_1, \dots, x_n]$ of polynomials in n commuting indeterminates inductively by

$$R[x_1, \dots, x_n, x_{n+1}] = (R[x_1, \dots, x_n])[x_{n+1}]$$

for each $n \geq 1$. Then iterating Theorem 10 gives

Corollary 1. *If R is a UFD, so also is $R[x_1, \dots, x_n]$ for each $n \geq 1$.*

Exercises 5.1

Throughout these exercises, R is an integral domain unless stated otherwise.

1. If $0 \neq a = bc$ in R , show that $a \sim b$ if and only if $c \sim 1$.
2. If $a \sim a'$ and $b \sim b'$ in R , show that $a|b$ if and only if $a'|b'$.
3. In the ring $\mathbb{Z}(i)$ of Gaussian integers, show that
 - (a) $(2+i) \sim (1-2i)$
 - (b) $(1+2i) \not\sim (2+i)$
4. Show that $(1-\sqrt{5}) \sim (7-3\sqrt{5})$ in $\mathbb{Z}(\sqrt{5})$.
5. If $a \sim a'$ and $b \sim b'$ in R , show that $ab \sim a'b'$.
6. Show that an integral domain is a field if and only if $a \sim b$ for all $a \neq 0 \neq b$.
7. Find the units in $\mathbb{Z}(\sqrt{-5})$. [Hint: Example 3.]
8. Find the units in $\mathbb{Z}(\sqrt{-3})$. [Hint: Use $N(a+b\sqrt{-3}) = a^2 + 3b^2$ as in Example 3.]

9. If R is an integral domain and $p \in R$, show that p is irreducible if and only if $\langle p \rangle \subseteq \langle a \rangle$, where $a \notin R^*$, implies that $\langle p \rangle = \langle a \rangle$.
10. In each case, determine whether p is irreducible in $\mathbb{Z}(i)$.
 - (a) $p = 11$
 - (b) $p = 2 - i$
 - (c) $p = 5$
 - (d) $p = 7 - i$
11. Let $p \in \mathbb{Z}$ be a prime and assume that $p \equiv 3 \pmod{4}$. Show that p is irreducible in $\mathbb{Z}(i)$. [Hint: Corollary to Theorem 8 §1.3.]
12. In each case, determine whether p is irreducible in $\mathbb{Z}(\sqrt{-5})$.
 - (a) $p = 6 + \sqrt{-5}$
 - (b) $p = 7$
 - (c) $p = 29$
 - (d) $p = 2 - 3\sqrt{-5}$
13. In each case show that p is irreducible in $\mathbb{Z}(\sqrt{-5})$ but is not a prime.
 - (a) $p = 2 + \sqrt{-5}$
 - (b) $p = 1 + 2\sqrt{-5}$
14. In each case, determine whether p is irreducible in $\mathbb{Z}(\sqrt{-3})$. [Hint: Use the norm function $N(m + n\sqrt{-3}) = m^2 + 3n^2$.]
 - (a) $p = 3 + 2\sqrt{-3}$
 - (b) $p = 2 + 3\sqrt{-3}$
 - (c) $p = 5$
 - (d) $p = 7$
15. Show that $1 + \sqrt{-3}$ is irreducible in $\mathbb{Z}(\sqrt{-3})$ but is not a prime.
16. Let $p \sim q$ in the integral domain R .
 - (a) Show that p is irreducible if and only if q is irreducible.
 - (b) Show that p is a prime if and only if q is a prime.
17. If $p \in \mathbb{Z}(\sqrt{-5})$, define $N(p)$ as in Example 3. If $N(p)$ is a prime in \mathbb{Z} , show that p is irreducible in $\mathbb{Z}(\sqrt{-5})$.
18. Show that $\mathbb{Z}(\sqrt{5})$ is not a UFD by showing that $1 + \sqrt{5}$ is an irreducible that is not a prime. [Hint: Use $N(m + n\sqrt{5}) = m^2 - 5n^2$.]
19. A commutative ring is said to satisfy the descending chain condition on principal ideals (DCCP) if $\langle a_1 \rangle \supseteq \langle a_2 \rangle \supseteq \dots$ in R implies that $a_n \sim a_{n+1} \sim \dots$ for some $n \geq 1$ (see Lemma 1). Show that an integral domain R satisfies the DCCP if and only if R is a field.
20. Prove $(2) \Rightarrow (1)$ in Lemma 1.
21. Show that R has ACCP if and only if any nonempty family F of principal ideals of R has a maximal member. [$\langle p \rangle$ in F is called maximal in F if $\langle p \rangle \subseteq \langle a \rangle$, with $\langle a \rangle$ in F , implies that $\langle p \rangle = \langle a \rangle$.]
22. Show that $F[x]$ satisfies the ACCP for any field F by modifying the argument in Example 4.
23. If S is a UFD and R is a subring, is R necessarily a UFD? Justify your answer.
24. Assume that $\gcd(a, b, \dots)$ and $\text{lcm}(a, b, \dots)$ exist in an integral domain.
 - (a) Show that $\gcd(0, a, b, \dots) \sim \gcd(a, b, \dots)$ and $\text{lcm}(0, a, b, \dots) \sim 0$.
 - (b) If u is a unit, show that $\text{lcm}(u, a, b, \dots) \sim \text{lcm}(a, b, \dots)$ and $\gcd(u, a, b, \dots) \sim 1$.
25. If $a_i \sim b_i$ in R for $i = 1, 2, \dots, n$, show that when they exist,
 - (a) $\gcd(a_1, \dots, a_n) \sim \gcd(b_1, \dots, b_n)$.
 - (b) $\text{lcm}(a_1, \dots, a_n) \sim \text{lcm}(b_1, \dots, b_n)$.
26. Show that $\gcd(ba_1, \dots, ba_n) \sim b \cdot \gcd(a_1, \dots, a_n)$ in R whenever both gcds exist in R and $b \neq 0$. This extends Lemma 4.
27. Show that $\gcd[a, \gcd(b, c)] \sim \gcd[\gcd(a, b), c]$ whenever all the gcds exist in R . Moreover, show that this common value is $\gcd(a, b, c)$.
28. If $\gcd(a, b) \sim 1 \sim \gcd(a, c)$ in a UFD, show that $\gcd(a, bc) \sim 1$.
29. In a UFD, if $a|bc$ and $\gcd(a, b) \sim 1$, show that $a|c$.
30. In a UFD, show that $\gcd(a, b) \text{lcm}(a, b) = ab$ for all $a \neq 0, b \neq 0$.

31. Show that $\text{lcm}(a_1, \dots, a_n)$ exists in an integral domain R if and only if the intersection $\langle a_1 \rangle \cap \dots \cap \langle a_n \rangle$ is a principal ideal.
32. Show that $\text{lcm}(da, db, dc, \dots) \sim d \text{lcm}(a, b, c, \dots)$ in a UFD for all nonzero d, a, b, c, \dots
33. Prove Lemma 5. [Hint: Exercise 26.]
34. Let R be a subring of an integral domain S such that (1) $R^* = S^*$, and (2) if $s \in S$ and $s|r, r \in R$, then $s \in R$. (For example, $S = R[x]$.)
 (a) Show that $p \in R$ is irreducible in R if and only if it is irreducible in S .
 (b) If S is a UFD, show that R is a UFD.
 (c) Prove the converse of Theorem 10: If $R[x]$ is a UFD, then R is a UFD.
35. Let R be a UFD and let $g|f$ in $R[x]$, where $f \neq 0$. If f is primitive, show that g is also primitive.
36. Show that an integral domain R is a UFD if and only if it satisfies the ACCP and $\text{lcm}(a, b)$ exists for all $a \neq 0, b \neq 0$ in R . [Hint: If $p|ab$, p irreducible, consider $m \sim \text{lcm}(a, p)$. Use the fact that $m|ap$ and $m|ab$.]
37. Let R be a UFD with field of quotients F , and let $f, g \in R[x]$. If f and g are primitive, show that $f \sim g$ in $R[x]$ if and only if $f \sim g$ in $F[x]$.
38. Let R be a UFD with field of quotients F . If $p \in R[x]$ is primitive, and p is irreducible in $F[x]$, show that p is irreducible in $R[x]$.
39. (P.M. Cohn) Fix a prime p in \mathbb{Z} and let R denote the set of all polynomials in $\mathbb{Z}[x]$ with the coefficient of x divisible by p .
 (a) Show that R is an integral domain.
 (b) Show that $\gcd(p, px) = 1$ in R but that $\text{lcm}(p, px)$ does not exist in R .
40. (T.W. Hungerford) Let R denote the set of polynomials f in $\mathbb{Q}[x]$ with constant coefficient in \mathbb{Z} .
 (a) Show that R is an integral domain, $R^* = \{1, -1\}$, and R is not a UFD. [Hint: Consider $x, \frac{1}{2}x, \frac{1}{4}x, \frac{1}{8}x, \dots$]
 (b) Show that $f \in R$ is irreducible if and only if f is one of two types: (1) $f \sim p$, p a prime in \mathbb{Z} ; (2) $f \sim h$, h irreducible in $\mathbb{Q}[x]$, of positive degree, with constant coefficient 1.
 (c) Show that each irreducible in R is prime.
 (d) Show that $\gcd(f, g)$ exists in R for all $f \neq 0, g \neq 0$ in R . [Hint: Consider whether $x|f$ or $x|g$.]
 (e) If $f \neq 0$ in R , show that $f = tx^n h_1 \cdots h_r$, where $t \in \mathbb{Q}$, $n \geq 0$, and each h_i is irreducible in \mathbb{Q} with constant coefficient 1. Moreover, if $f = t'x^{n'} h'_1 \cdots h'_s$ is another such representation, show that $t = t'$, $n = n'$, $r = s$, and (after relabeling) $h_i = \pm h'_i$ for all i .

5.2 PRINCIPAL IDEAL DOMAINS

Theorem 3 §5.1 shows that an integral domain satisfying the ascending chain condition on principal ideals has the property that every nonzero nonunit factors into irreducibles. It turns out that the following family of integral domains all have this property.

An integral domain R is called a **principal ideal domain (PID)** if every ideal is principal.

Example 1. \mathbb{Z} is a PID. Indeed, every additive subgroup of \mathbb{Z} is cyclic of the form $\langle m \rangle = m\mathbb{Z}$ for some $m \in \mathbb{Z}$.

Example 2. If F is a field, $F[x]$ is a PID by Theorem 1 §4.3.

Principal ideal domains are quite common. For example, every ring R such that $\mathbb{Z} \subseteq R \subseteq \mathbb{Q}$ is a PID (see Exercise 10). In addition, we will also show that the ring $\mathbb{Z}(i)$ of gaussian integers is a PID.

One of the most useful facts about \mathbb{Z} (and about $F[x]$, where F is a field) is that any two elements have a greatest common divisor that is a linear combination of them. This is true in every PID.

Theorem 1. Let R be a PID, and let a_1, a_2, \dots, a_n be nonzero elements of R . Then $d \sim \gcd(a_1, \dots, a_n)$ exists, and there exist $r_1, \dots, r_n \in R$ such that

$$\gcd(a_1, \dots, a_n) = r_1 a_1 + \dots + r_n a_n.$$

Proof. Let $A = \{r_1 a_1 + \dots + r_n a_n \mid r_i \in R\}$ denote the set of all linear combinations of the a_i . Then A is an ideal of R as is easily verified, so $A = \langle d \rangle$ for some $d \in R$ because R is a PID. Thus, $d = r_1 a_1 + \dots + r_n a_n$ for some r_i , and we claim that $d \sim \gcd(a_1, \dots, a_n)$. We have $d|a_i$ for each i because $a_i \in A$. But if $r|a_i$ for all i , then r clearly divides $d = r_1 a_1 + r_2 a_2 + \dots + r_n a_n$. Thus, $d \sim \gcd(a_1, \dots, a_n)$ by the definition of the gcd, which proves Theorem 1. ■

The next theorem reconfirms the fact that \mathbb{Z} and $F[x]$ are UFDs.

Theorem 2. Every PID is a UFD.

Proof. If R is a PID, it suffices to verify the ACCP by Theorem 1 and Theorem 7 §5.1. If $\langle a_1 \rangle \subseteq \langle a_2 \rangle \subseteq \dots$ in R , put $A = \langle a_1 \rangle \cup \langle a_2 \rangle \cup \dots$. Then A is an ideal (verify), so let $A = \langle a \rangle$ by hypothesis. Thus, $a \in \langle a_n \rangle$ for some n , so

$$\langle a \rangle \subseteq \langle a_n \rangle \subseteq \langle a_{n+1} \rangle \subseteq \dots \subseteq A = \langle a \rangle.$$

Hence, $\langle a_n \rangle = \langle a_{n+1} \rangle = \dots$, as required by Lemma 1 §5.1. ■

The converse of Theorem 2 is false as the next example shows.

Example 3. Show that $\mathbb{Z}[x]$ is a UFD that is not a PID.

Solution. $\mathbb{Z}[x]$ is a UFD by Theorem 10 §5.1. Let $A = \{2n + xf \mid n \in \mathbb{Z}, f \in \mathbb{Z}[x]\}$. Then A is an ideal of $\mathbb{Z}[x]$, which we claim is not principal. For if $A = \langle g \rangle$, $g \in \mathbb{Z}[x]$, then $g|2$ because $2 \in A$, so $g = \pm 1$ or $g = \pm 2$. But then $A = \mathbb{Z}[x]$ or $A = \langle 2 \rangle$, respectively, and both these possibilities are false. □

If R is a UFD, the least common multiple exists for any finite set of nonzero elements of R (Theorem 6 §5.1). In a PID we can say more. If A_1, A_2, \dots, A_n are ideals of any ring R , their sum is defined by

$$A_1 + A_2 + \dots + A_n = \{x_1 + x_2 + \dots + x_n \mid x_i \in A_i \text{ for all } i\}.$$

This is easily verified to be an ideal of R containing each A_i . In particular, if a_1, a_2, \dots, a_n are nonzero elements of R , then

$$\langle a_1 \rangle + \langle a_2 \rangle + \dots + \langle a_n \rangle = \{r_1 a_1 + r_2 a_2 + \dots + r_n a_n \mid r_i \in R\}$$

is the ideal considered in the proof of Theorem 1. Hence, that proof shows that (1) below is true. The *dual* holds too:

- $$\begin{aligned} (1) \quad d &\sim \gcd(a_1, \dots, a_n) & \text{if and only if} & \langle a_1 \rangle + \langle a_2 \rangle + \dots + \langle a_n \rangle = \langle d \rangle. \\ (2) \quad m &\sim \operatorname{lcm}(a_1, \dots, a_n) & \text{if and only if} & \langle a_1 \rangle \cap \langle a_2 \rangle \cap \dots \cap \langle a_n \rangle = \langle m \rangle. \end{aligned}$$

We leave verification of (2) to the reader (see Exercise 31 §5.1).

The fact that \mathbb{Z}_p is a field whenever $p \in \mathbb{Z}$ is a prime has the following analogue in an arbitrary PID.

Theorem 3. *The following are equivalent for a nonzero nonunit p in a PID R :*

- (1) p is a prime.
- (2) $R/\langle p \rangle$ is a field.
- (3) $R/\langle p \rangle$ is an integral domain.

In particular, every nonzero prime ideal of R is maximal.

Proof. (1) \Rightarrow (2). Consider $x = a + \langle p \rangle$ in $R/\langle p \rangle$ and assume that $x \neq 0$. Then $a \notin \langle p \rangle$; that is, p does not divide a . Now $A = \{ra + sp \mid r, s \in R\}$ is an ideal of R , so, as R is a PID, let $A = \langle d \rangle$, $d \in R$. Then $d|p$ because $p \in A$, so $d \sim p$ or $d \sim 1$. But $d \sim p$ means that $\langle p \rangle = \langle d \rangle = A$, whence $p|a$ because $a \in A$, contrary to assumption. So $d \sim 1$, from which $A = \langle 1 \rangle = R$. In particular, $1 \in A$, say $1 = ba + sp$. If $y = b + \langle p \rangle$, then $xy = 1$ in $R/\langle p \rangle$, proving (2).

(2) \Rightarrow (3). Every field is an integral domain.

(3) \Rightarrow (1). Suppose that $p|ab$ in R ; we must show that $p|a$ or $p|b$. Now $p|ab$ implies that $(a + \langle p \rangle)(b + \langle p \rangle) = 0$ in $R/\langle p \rangle$ so, by (3), $a + \langle p \rangle = 0$ or $b + \langle p \rangle = 0$ in $R/\langle p \rangle$. Thus $p|a$ or $p|b$, proving (1).

Finally, let A be a prime ideal of R , say $A = \langle p \rangle$, $p \in R$. Then R/A is an integral domain, hence a field because (3) \Rightarrow (2). Thus, A is a maximal ideal of R by Corollary 1 of Theorem 6 §3.3, proving the last sentence of the theorem. ■

Theorem 3 shows that nonzero prime ideals in a PID are maximal. However, this property may fail in a UFD. For example, if $R = \mathbb{Z}[x]$, then R is a UFD by Theorem 10 §5.1. But $\langle x \rangle$ is a prime ideal of $\mathbb{Z}[x]$ that is not maximal in $\mathbb{Z}[x]$, because $R/\langle x \rangle \cong \mathbb{Z}$ is an integral domain that is not a field.

Division Algorithms

Another useful property of the integral domains \mathbb{Z} and $F[x]$, F a field, is that both possess a division algorithm. This property leads to an interesting class of PIDs.

In general, if R is an integral domain, we say that R has a **division algorithm** if a map $\delta: R \rightarrow \mathbb{N}$ exists (called a **divisor function**) such that the following condition is satisfied:

DA Given a and $b \neq 0$ in R , there exist q and r in R such that
 $a = qb + r$ and either $r = 0$ or $\delta(r) < \delta(b)$.

Example 4. If F is a field, $\delta(f) = \deg f$ is a divisor function for $F[x]$ by Theorem 4 §4.1.

Example 5. Show that $\delta(a) = |a|$ is a divisor function for \mathbb{Z} .

Solution. Let $a, b \in \mathbb{Z}$ where $b \neq 0$. Then $|b| > 0$, so $a = q|b| + r$ by Theorem 1 §1.2, where $0 \leq r < |b|$. If $b > 0$, this reads $a = qb + r$; if $b < 0$, it becomes $a = q(-b) + r = (-q)b + r$. Hence, DA holds in both cases. \square

Theorem 4. Every integral domain R with a division algorithm is a PID.

Proof. Let R be such a domain with divisor function δ , and let B be an ideal of R . If $B = 0$, then $B = \langle 0 \rangle$ is principal. Otherwise, let $0 \neq b \in B$ be such that $\delta(b)$ is as small as possible. Thus, $\langle b \rangle \subseteq B$ and we claim this is equality. Given $a \in B$, write $a = qb + r$, where $r = 0$ or $\delta(r) < \delta(b)$. But $r = a - qb \in B$, so this is a contradiction if $r \neq 0$. Hence, $r = 0$, and so $a = qb \in \langle b \rangle$. It follows that $B \subseteq \langle b \rangle$, as required. ■

Let R be an integral domain with a division algorithm. If a and b are nonzero elements of R , we can compute $\gcd(a, b)$ using the Euclidean algorithm. Since the procedure is entirely analogous to that in \mathbb{Z} (Section 1.2), we merely sketch it here. The idea is to use DA repeatedly as follows:

$$\begin{array}{lllll} a = q_1b + r_1, & \text{where } & r_1 = 0 & \text{or} & \delta(r_1) < \delta(b), \\ b = q_2r_1 + r_2, & \text{where } & r_2 = 0 & \text{or} & \delta(r_2) < \delta(r_1), \\ r_1 = q_3r_2 + r_3, & \text{where } & r_3 = 0 & \text{or} & \delta(r_3) < \delta(r_2), \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{m-1} = q_{m+1}r_m + r_{m+1}, & \text{where } & r_{m+1} = 0 & \text{or} & \delta(r_{m+1}) < \delta(r_m), \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{array}$$

Because $\delta(b) > \delta(r_1) > \delta(r_2) > \dots$ is a sequence of nonnegative integers, the process must encounter $r_{m+1} = 0$ at some stage where $r_m \neq 0$. Then $r_m|r_{m-1}$, so $\gcd(r_{m-1}, r_m) \sim r_m$. Now, as for \mathbb{Z} (see Example 4 §1.2), we get

$$\gcd(a, b) \sim \gcd(b, r_1) \sim \gcd(r_1, r_2) \sim \dots \sim \gcd(r_{m-1}, r_m) \sim r_m.$$

Thus, the algorithm produces $\gcd(a, b)$ as the last nonzero remainder. Finally, just as for \mathbb{Z} , elimination of remainders in the preceding equations gives $\gcd(a, b) = r_m$ in the form $r_m = ra + sb$ with $r, s \in R$, as guaranteed by Theorem 1.

Some Rings of Quadratic Integers

We are going to show that the ring $\mathbb{Z}(i)$ of Gaussian integers has a division algorithm. The method we use applies to other integral domains that resemble $\mathbb{Z}(i)$. For instance, Example 5 §5.1 shows that the integral domain $\mathbb{Z}(\sqrt{-5})$ is not a UFD because it contains an element $p = 1 + \sqrt{-5}$, which is irreducible but not prime. The study of such subrings of \mathbb{C} was the source of the mathematics presented in this chapter, and it has now evolved into a subject in its own right, algebraic number theory. Note that i and $\sqrt{-5}$ are roots of quadratic polynomials $x^2 + 1$ and $x^2 + 5$, respectively. Accordingly, we discuss subrings of \mathbb{C} that result from adjoining a root of some quadratic $x^2 + m$ to \mathbb{Z} , where $m \in \mathbb{C}$.

Throughout the discussion, ω denotes a complex number such that

$$\omega^2 \in \mathbb{Z} \quad \text{and} \quad \omega \notin \mathbb{Q}.$$

The ring selected for study is

$$\mathbb{Z}(\omega) = \{m + n\omega \mid m, n \in \mathbb{Z}\}.$$

Clearly, $\mathbb{Z}(\omega)$ is a subring of \mathbb{C} and so is an integral domain. Moreover, the representation $m + n\omega$ of elements of $\mathbb{Z}(\omega)$ is unique in the following sense:

$$\text{If } m + n\omega = m' + n'\omega \text{ in } \mathbb{Z}(\omega) \quad \text{then} \quad m = m' \text{ and } n = n'.$$

Indeed, if $m + n\omega = m' + n'\omega$, then $(n - n')\omega = m' - m$. If $n' \neq n$, this gives $\omega \in \mathbb{Q}$, contrary to our assumption. Hence, $n' = n$ and so $m' = m$.

Most of what we have to say about $\mathbb{Z}(\omega)$ depends on two fundamental notions. If $a = m + n\omega \in \mathbb{Z}(\omega)$, define the **conjugate** a^* of a and the **norm** $N(a)$ of a by

$$a^* = m - n\omega \quad \text{and} \quad N(a) = m^2 - \omega^2 n^2.$$

Thus $a^* \in \mathbb{Z}(\omega)$, and $N(a) \in \mathbb{Z}$ for all $a \in \mathbb{Z}(\omega)$ because $\omega^2 \in \mathbb{Z}$.

Example 6. If $a = m + ni$ in $\mathbb{Z}(i)$, then $a^* = m - ni$ is the usual complex conjugate of a , and $N(a) = m^2 + n^2 = |a|^2$ is the square of the usual absolute value of a .

Example 7. If $a = m + n\sqrt{-5}$ in $\mathbb{Z}(\sqrt{-5})$, then the conjugate is $a^* = m - n\sqrt{-5}$ and $N(a) = m^2 + 5n^2$. This coincides with the usage in Example 3 §5.1.

The next theorem collects several basic properties of norms and conjugates in $\mathbb{Z}(\omega)$. In the case of the gaussian integers $\mathbb{Z}(i)$, these properties reduce to familiar facts about the complex numbers.

Theorem 5. Let $\omega \in \mathbb{C}$ satisfy $\omega^2 \in \mathbb{Z}$, $\omega \notin \mathbb{Q}$. Then the following properties hold for all a and b in $\mathbb{Z}(\omega)$.

- (1) $aa^* = N(a) = N(a^*)$.
- (2) $(ab)^* = a^*b^*$ and $a^{**} = a$.
- (3) $N(ab) = N(a)N(b)$.
- (4) a is a unit in $\mathbb{Z}(\omega)$ if and only if $N(a) = \pm 1$, and then $a^{-1} = N(a)a^*$.
- (5) $N(a) = 0$ if and only if $a = 0$.
- (6) If $N(a)$ is a prime in \mathbb{Z} , then a is irreducible in $\mathbb{Z}(\omega)$.

Proof. (1) and (2). The routine verifications are left as Exercise 11.

(3). By (1) and (2), $N(ab) = (ab)(ab)^* = aba^*b^* = aa^*bb^* = N(a)N(b)$.

(4). If a is a unit, then (3) gives $N(a)N(a^{-1}) = N(1) = 1$, so $N(a) = \pm 1$. Conversely, if $N(a) = \pm 1$, then $a[N(a)a^*] = N(a)^2 = 1$ by (1). Thus, $a^{-1} = N(a)a^*$.

(5). If $a = m + n\omega$, then $N(a) = 0$ means that $m^2 - \omega^2 n^2 = 0$. If $n \neq 0$, this gives $\omega = \pm(m/n) \in \mathbb{Q}$, contrary to assumption. So $n = 0$, from which $m = 0$, and hence $a = 0$. The converse is clear.

(6). If $N(a)$ is a prime in \mathbb{Z} , let $a = bc$ in $\mathbb{Z}(\omega)$. Then $N(a) = N(b)N(c)$ in \mathbb{Z} , so $N(b) = \pm 1$ or $N(c) = \pm 1$. Hence, b or c is a unit in $\mathbb{Z}(\omega)$ by (4). ■

Note that the converse to (6) of Theorem 5 is not true: In $\mathbb{Z}(\sqrt{-5})$, the element $a = 1 + \sqrt{-5}$ is irreducible (Example 5 §5.1) but $N(a) = 6$ is not prime.

Of course, the units in $\mathbb{Z}(\omega)$ are of interest. For example, if $a = m + n\sqrt{-2}$ is a unit in $\mathbb{Z}(\sqrt{-2})$, then $m^2 + 2n^2 = N(a) = \pm 1$ by Theorem 5. This easily shows that 1 and -1 are the only units in $\mathbb{Z}(\sqrt{-2})$. In fact, this holds for $\mathbb{Z}(\sqrt{-d})$, where $d > 0$ is any integer that is not a square.

On the other hand, $a = m + n\sqrt{2}$ is a unit in $\mathbb{Z}(\sqrt{2})$ if and only if the norm $N(a) = m^2 - 2n^2 = \pm 1$. In particular, $u = 1 + \sqrt{2}$ is a unit in $\mathbb{Z}(\sqrt{2})$ where

$u^{-1} = -1 + \sqrt{2}$. Hence, $\pm u^k$ is a unit for any $k \in \mathbb{Z}$. (In fact, these are *all* the units in $\mathbb{Z}(\sqrt{2})$; see Exercise 31.) In this case, if $d > 0$ is any integer that is not a square, then $m + n\sqrt{d}$ is a unit in $\mathbb{Z}(\sqrt{d})$ if and only if

$$m^2 - dn^2 = \pm 1.$$

This is sometimes called **Pell's equation**, and solutions with $m \neq \pm 1$ always exist.⁶⁹ Hence, $\mathbb{Z}(\sqrt{d})$ has a unit $u \neq \pm 1$, so taking powers of u gives infinitely many solutions of Pell's equation, an observation made originally by Fermat.

Example 8. If $d > 0$ is a nonsquare integer, then $\mathbb{Z}(\sqrt{d})$ has infinitely many units.

We now turn to the factorization theory in $\mathbb{Z}(\omega)$.

Theorem 6. Every nonzero nonunit in $\mathbb{Z}(\omega)$ is a product of irreducibles.

Proof. By Theorem 3 §5.1, it suffices to show that $\mathbb{Z}(\omega)$ satisfies the ACCP; that is, a strictly increasing chain $0 \subset \langle a_1 \rangle \subset \langle a_2 \rangle \subset \dots$ in $\mathbb{Z}(\omega)$ is impossible. Suppose such a chain exists. Then $a_{n+1}|a_n$ for each $n \geq 1$, say $a_n = b_n a_{n+1}$. Moreover, b_n is not a unit because $\langle a_n \rangle \neq \langle a_{n+1} \rangle$, so $|N(b_n)| > 1$ by Theorem 5. But then $|N(a_n)| > |N(a_{n+1})|$ because $N(a_n) = N(b_n)N(a_{n+1})$, so $|N(a_1)| > |N(a_2)| > \dots$ is a strictly decreasing sequence of nonnegative integers, a contradiction. ■

Theorem 6 notwithstanding, it is difficult to determine which choices of ω make $\mathbb{Z}(\omega)$ into a UFD or a PID. We content ourselves with one condition that guarantees that $\mathbb{Z}(\omega)$ has a division algorithm. We need a technical lemma.

Lemma 1. Assume that, for any r and s in \mathbb{Q} , there exist m and n in \mathbb{Z} such that

$$|(r - m)^2 - \omega^2(s - n)^2| < 1.$$

Then $\delta(a) = |N(a)|$ defines a divisor function $\mathbb{Z}(\omega) \rightarrow \mathbb{N}$.

Proof. Given a and $b \neq 0$ in $\mathbb{Z}(\omega)$, we must find r and q in $\mathbb{Z}(\omega)$ such that $a = qb + r$ and either $r = 0$ or $\delta(r) < \delta(b)$. Working in \mathbb{C} yields

$$\frac{a}{b} = \frac{ab^*}{bb^*} = \frac{ab^*}{N(b)}.$$

Hence, a/b has the form $a/b = r + sw$, where r and s are in \mathbb{Q} , so $a = (r + sw)b$. Choose m and n as in the hypothesis, write $q = m + n\omega$, and define

$$r = -qb + a = (-m - n\omega)b + (r + sw)b = [(r - m) + (s - n)\omega]b.$$

Now observe that $\delta(xy) = \delta(x)\delta(y)$ for all $x, y \in \mathbb{Z}(\omega)$ by Theorem 5. Then

$$\delta(r) = \delta[(r - m) + (s - n)\omega]\delta(b) = |(r - m)^2 - \omega^2(s - n)^2|\delta(b) < \delta(b),$$

which proves DA. □

Theorem 7. The ring $\mathbb{Z}(i)$ of gaussian integers has a division algorithm with divisor function $\delta(a) = |N(a)|$ for all $a \in \mathbb{Z}(i)$. That is $\delta(m + ni) = m^2 + n^2$ for all $m, n \in \mathbb{Z}$.

Proof. If r is a rational number and m is the integer closest to r , it is clear that $|r - m| \leq \frac{1}{2}$. Similarly, let $|s - n| \leq \frac{1}{2}$, $n \in \mathbb{Z}$. Thus, Lemma 1 applies because

$$(r - m)^2 - i^2(s - n)^2 = (r - m)^2 + (s - n)^2 \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

□

⁶⁹For a detailed discussion using continued fractions, see Davenport, H., *The Higher Arithmetic*, New York: Harper, 1960.

Euclidean Domains

In Theorem 7, the division function $\delta(a) = |N(a)|$ on $\mathbb{Z}(i)$ has the property that $\delta(ab) = \delta(a)\delta(b)$ for all $a, b \in \mathbb{Z}(i)$. Since $\delta(a) = 0$ if and only if $a = 0$ by Theorem 5, it follows that $\delta(a) \geq 1$ whenever $a \neq 0$, and hence $\delta(ab) \geq \delta(a)$ whenever $a \neq 0$ and $b \neq 0$. This suggests the following definition:

An integral domain R is called a **euclidean domain** if it has a division algorithm with divisor function $\delta : R \rightarrow \mathbb{N}$ that satisfies DA and the following condition:

$$E \quad \text{If } a \neq 0 \text{ and } b \neq 0 \text{ in } R, \text{ then } \delta(ab) \geq \delta(a).$$

Example 9. The ring \mathbb{Z} is euclidean where $\delta(a) = |a|$ for all $a \neq 0$ in \mathbb{Z} .

Example 10. If F is a field, $F[x]$ is euclidean if $\delta(f) = \deg f$ for all $f \neq 0$ in $F[x]$.

Example 11. The gaussian integers $\mathbb{Z}(i)$ is euclidean if $\delta(m + ni) = \sqrt{m^2 + n^2}$.

Some PIDs are not euclidean, but examples are not easy to find. In 1949, T. Motzkin provided the first such example: $\mathbb{Z}(\frac{1}{2}(1 + \sqrt{-19}))$ is a non-euclidean PID.⁷⁰ The extra condition E gives more information about the euclidean ring in terms of the mapping δ . Example 12 characterizes the units in terms of δ .

Example 12. Let R be a euclidean domain. If $0 \neq a \in R$, show that $\delta(1) \leq \delta(a)$ and that a is a unit if and only if $\delta(1) = \delta(a)$.

Solution. We have $\delta(1) \leq \delta(1 \cdot a) = \delta(a)$ by E. If a is a unit, $\delta(a) \leq \delta(aa^{-1}) = \delta(1)$, again by E, so $\delta(a) = \delta(1)$. Conversely, if $\delta(a) = \delta(1)$, write $1 = qa + r$, where $r = 0$ or $\delta(r) < \delta(a)$. If $r \neq 0$, we obtain $\delta(1) \leq \delta(1 \cdot r) < \delta(a) = \delta(1)$, a contradiction. So $r = 0$, $1 = qa$, and a is a unit. \square

There is a lot of information available on these euclidean domains, but we will not pursue this here. Instead, given a and $b \neq 0$ in $\mathbb{Z}(i)$, we conclude by giving an example of how the technique used to prove Lemma 1 can be employed to actually find q and r in $\mathbb{Z}(i)$ such that $a = qb + r$ and either $r = 0$ or $\delta(r) < \delta(b)$.

Example 13. Let $a = 7 + 8i$ and $b = 2 - i$ in $\mathbb{Z}(i)$. Find q and r in $\mathbb{Z}(i)$ such that $a = qb + r$ and either $r = 0$ or $\delta(r) < \delta(b)$.

Solution. The technique in Lemma 1 applies. Compute in \mathbb{C}

$$\frac{a}{b} = \frac{a\bar{b}}{b\bar{b}} = \frac{(7+8i)(2+i)}{2^2+1^2} = \frac{6+23i}{5}.$$

Now the closest integers to $6/5$ and $23/5$ are 1 and 5 , respectively. Hence, we write $6/5 = 1 + 1/5$ and $23/5 = 5 - 2/5$ to get

$$\frac{a}{b} = \frac{6}{5} + \frac{23}{5}i = (1 + 5i) + \left(\frac{1}{5} - \frac{2}{5}i\right).$$

⁷⁰See Motzkin, T., Bulletin of the American Mathematical Society, 55 (1949), 1142–1146. See also Càmpoli, O.A., *American Mathematical Monthly*, 95 (1988), 868–871; Wilson, J.C., *Mathematics Magazine*, 46 (1973), 34–38; Williams, K.S., *Mathematics Magazine*, 48 (1975), 176–177.

Thus

$$a = (1 + 5i)b + \left(\frac{1}{5} - \frac{2}{5}i\right)b = (1 + 5i)b + (0 - i),$$

so $q = 1 + 5i$ and $r = -i$. Note that $\delta(r) = 1 < 5 = \delta(b)$. \square

Ernst Eduard Kummer (1810–1893) Kummer entered the University of Halle at the age of 18 and within 3 years had a Ph.D. in mathematics. He became a professor at the University of Breslau in 1842, and in 1855 he succeeded Dirichlet at the University of Berlin. Kummer is best remembered as the creator, with Dedekind and Kronecker, of algebraic number theory. As described in the introduction to this chapter, Kummer was interested in Fermat's last theorem and was led to consider why the unique factorization into primes failed in $\mathbb{Z}(\omega)$, where ω is a root of unity. His creation of ideal numbers, for which the uniqueness can be restored, has been compared to the creation of non-Euclidean geometry. Its importance as a mathematical achievement stems from the fact that it led, via Dedekind, to the modern notion of an ideal.

In addition to algebra, Kummer also made contributions to geometry, analysis, and physics. He was a popular lecturer and directed many Ph.D. students. In 1857, he was awarded the grand prize in mathematics of the French Academy of Sciences.

Exercises 5.2

1. Is every subring of a PID again a PID? Support your answer.
2. If F is a field, show that $F[x, y]$ is a UFD that is not a PID. [Hint: Consider $\{f \mid f(0, 0) = 0\}$.]
3. Show that every field F is a PID.
4. Is $\mathbb{Z}(\sqrt{-5})$ a PID? Defend your answer.
5. If R is a PID and $A \neq 0$ is an ideal of R , show that R/A has a finite number of ideals, all of which are principal.
6. (a) Is every prime ideal of a PID maximal? Support your answer.
 (b) Show that every ideal $A \neq R$ in a PID R is contained in a maximal ideal of R .
7. Show that the following conditions are equivalent for an integral domain R .
 (a) R is a field, (b) $R[x]$ is Euclidean, and (c) $R[x]$ is a PID.
8. Let $p \in \mathbb{Z}$ be a prime and define $\mathbb{Z}_{(p)} = \left\{ \frac{m}{n} \in \mathbb{Q} \mid p \text{ does not divide } n \right\}$.
 (a) Show that $\mathbb{Z}_{(p)}$ is an integral domain (called the **localization** of \mathbb{Z} at p) and find the units.
 (b) If $A \neq 0$ is an ideal of $\mathbb{Z}_{(p)}$, show that $A = \langle p^k \rangle$, where $k \geq 0$ is the smallest integer such that $p^k \in A$. [Hint: If $0 \neq m \in \mathbb{Z}$, then $m = p^r d$, where $r \geq 0$ and p does not divide d .]
 (c) Show that $\mathbb{Z}_{(p)}$ is a PID with exactly one maximal ideal.
9. Let $\mathbb{Z}_{(p)}$ be as in Exercise 8. Show that $\mathbb{Z}_{(p)}$ is a Euclidean domain where, for each $a \neq 0$ in R , $\delta(a) = k$ where $\langle a \rangle = \langle p^k \rangle$. Indeed, show that $\delta(ab) = \delta(a) + \delta(b)$ for all $a \neq 0, b \neq 0$ in $\mathbb{Z}_{(p)}$ and that, if $a + b \neq 0$, then $\delta(a + b) \geq \min\{\delta(a), \delta(b)\}$.
10. Let R be a ring such that $\mathbb{Z} \subseteq R \subseteq \mathbb{Q}$. Show that R is a PID. [Hint: If I is an ideal of R , consider $A = \mathbb{Z} \cap I$.]
11. (a) Prove (1) and (2) of Theorem 5.
 (b) Prove that the converse of (6) in Theorem 5 is false. [Hint: In Example 5 §5.1, consider $a = 1 + \sqrt{-5}$.]

12. Let ω be as in Theorem 5 and assume that $\omega^2 < 0$. Show that $\mathbb{Z}(\omega)$ has finitely many units.
13. (a) Show that $\mathbb{Z}(\sqrt{-2})$ is euclidean with $\delta(a) = |N(a)|$.
 (b) If $a = 4 + 3\sqrt{-2}$ and $b = 3 - \sqrt{-2}$, write $a = qb + r$, where $r = 0$ or $\delta(r) < \delta(b)$.
14. (a) Show that $\mathbb{Z}(\sqrt{2})$ is euclidean with $\delta(a) = |N(a)|$.
 (b) If $a = 5 + 7\sqrt{2}$ and $b = 3 + \sqrt{2}$, write $a = qb + r$, where $r = 0$ or $\delta(r) < \delta(b)$.
15. (a) Show that $\mathbb{Z}(\sqrt{3})$ is euclidean with $\delta(a) = |N(a)|$.
 (b) If $a = 4 + 5\sqrt{3}$ and $b = 1 + \sqrt{3}$, write $a = qb + r$, where $r = 0$ or $\delta(r) < \delta(b)$.
16. Show that $\mathbb{Z}(\sqrt{-3})$ is not euclidean with $\delta(a) = |N(a)|$. [Hint: Try $a = 1 + \sqrt{-3}$ and $b = 2$.]
17. If R is a euclidean domain, and if $m > 0$ and k are integers, show that δ' satisfies DA and E, where $\delta'(a) = m \cdot \delta(a) + k$.
18. (a) If F is a field, show that F is euclidean.
 (b) If the mapping δ is constant in a euclidean domain R , show that R is a field.
19. If $a \sim b$ in a euclidean domain R , show that $\delta(a) = \delta(b)$.
20. If $a|b$ and $\delta(a) = \delta(b)$ in a euclidean domain R , show that $a \sim b$.
21. Let $b \neq 0$ in a euclidean domain R . Show that b is a nonunit if and only if $\delta(ab) > \delta(a)$ for all $a \neq 0$ in R . [Hint: Exercises 19 and 20.]
22. Assume that R is a euclidean domain in which $\delta(a+b) \leq \max\{\delta(a), \delta(b)\}$ whenever a, b , and $a+b$ are nonzero. Show that q and r are uniquely determined in DA.
23. Suppose that a euclidean domain R has a unique maximal ideal P . Write $P = \langle p \rangle$ by Theorem 4.
 (a) Show that P consists of nonunits; that is, a is a nonunit if and only if $p|a$. [Hint: Exercise 6.]
 (b) Show that every ideal $A \neq 0$ of R has the form $A = \langle p^k \rangle$ for some $k \geq 0$.
24. (a) If $A = \langle 1+i \rangle$ in $\mathbb{Z}(i)$, show that $\mathbb{Z}(i)/A$ is a finite field and find its order.
 (b) If $A = \langle 1+2i \rangle$ in $\mathbb{Z}(i)$, show that $\mathbb{Z}(i)/A$ is a finite field and find its order.
25. For ω as in Theorem 5, show that $\mathbb{Q}(\omega) = \{r + sw \mid r, s \in \mathbb{Q}\}$ is the field of quotients of $\mathbb{Z}(\omega)$.
26. An ideal A of a commutative ring R is said to be **finitely generated** if we have $A = \{r_1a_1 + r_2a_2 + \cdots + r_na_n \mid r_i \in R\}$ for some a_1, a_2, \dots, a_n in A . We write $A = \langle a_1, \dots, a_n \rangle$ in this case and say that a_1, a_2, \dots, a_n generate A .
 (a) Show that the following conditions are equivalent for an integral domain R (then called a **Bézout domain**):
 (1) Every 2-generated ideal $A = \langle a, b \rangle$ is principal.
 (2) If $a \neq 0$ and $b \neq 0$, then $d = \gcd(a, b)$ exists and $d = ra + sb$ for some $r, s \in R$.
 (b) If R is a Bézout domain, show that every finitely generated ideal is principal; in fact, for all a_1, \dots, a_n in R , show that $d \sim \gcd(a_1, \dots, a_n)$ exists and that $\langle a_1, \dots, a_n \rangle = \langle d \rangle$.
27. Let R be an integral domain. Show that R is a PID if and only if it satisfies the ACCP and each 2-generated ideal $\langle a, b \rangle$ is principal. [Hint: Exercise 26.]
28. Let R be a UFD. Show that R is a PID if and only if for all $a \neq 0$ and $b \neq 0$ in R , r and s exist in R such that $\gcd(a, b) \sim ra + sb$. [Hint: Exercises 26 and 27.]
29. Let $a = bc$ in a PID R where $\gcd(b, c) \sim 1$. Show that $\frac{R}{(a)} \cong \frac{R}{(b)} \times \frac{R}{(c)}$. [Hint: Chinese remainder theorem, Theorem 8 §3.4.]

30. Let R be a PID and let A be an ideal of R that satisfies the condition that $r^2 \in A$, $r \in R$, implies that $r \in A$. Show that R/A is isomorphic to a finite direct product of fields. [Hint: Exercise 44 §3.4.]

31. Show that every unit of $\mathbb{Z}(\sqrt{2})$ has the form $\pm u^k$, where $k \in \mathbb{Z}$ and $u = 1 + \sqrt{2}$. [Hint: If $v > 0$ is a unit in $\mathbb{Z}(\sqrt{2})$, show that either $v = u^k$ for some integer k or $u^k < v < u^{k+1}$ for some k . Rule out the second case by showing that $1 < v < u$ is impossible if v is a unit ($v > 1$ implies that $-1 < v^* < 1$).]

32. For $a = m + nw$ in $\mathbb{Z}(\omega)$, define the **integral part** of a by $\text{int } a = m$. Then write $\langle a, b \rangle = \text{int}(ab^*)$ for all a and b in $\mathbb{Z}(\omega)$. If ω is as in Theorem 5, prove that the following hold for all a, b , and c in $\mathbb{Z}(\omega)$.

 - (a) $\langle a, b \rangle = \langle b, a \rangle$
 - (c) $\langle a + b, c \rangle = \langle a, c \rangle + \langle b, c \rangle$
 - (b) $\langle ka, b \rangle = k \langle a, b \rangle$ for all $k \in \mathbb{Z}$
 - (d) $\langle a, a \rangle = N(a)$

33. Can the integral domain $\mathbb{Z}(\sqrt{-2})$ be ordered (Section 3.5)? Defend your answer.

34. (a) Show that $\theta : \mathbb{Z}(\omega) \rightarrow M_2(\mathbb{Z})$ is a one-to-one ring homomorphism if

$$\theta(m + nw) = \begin{bmatrix} m & nw^2 \\ n & m \end{bmatrix}.$$

(b) Show that $N(a) = \det[\theta(a)]$ for all $a \in \mathbb{Z}(\omega)$.

35. Let $R = \mathbb{Z}(\omega)$, where ω is as in Theorem 5, and define $\tau : R \rightarrow R$ by $\tau(a) = a^*$ for all $a \in R$.

 - (a) Show that τ is a ring automorphism satisfying $\tau^2 = 1_R$.
 - (b) If $\sigma : R \rightarrow R$ is a ring automorphism satisfying $\sigma^2 = 1_R$, show that $\sigma = \tau$ or $\sigma = 1_R$.

36. If R is a PID and $A \neq 0$ is an ideal of R , show that every ideal of R/A is the annihilator of an element. [Hint: Every ideal of R/A has the form B/A . If $A = \langle a \rangle$, $B = \langle b \rangle$, then $a = bc$, $c \in R$. Show that $B/A = \text{ann}(c + A)$.]

Chapter 6

Fields

There is astonishing imagination, even in the science of mathematics.... We repeat, there was more imagination in the head of Archimedes than in that of Homer.

—Voltaire

Human beings have sought solutions to algebraic equations for centuries. This search has inspired some of the most creative (and important) mathematics imaginable. Suppose that a primitive tribe, motivated by the desire to count things and to tell others the results, has developed a facility with the set $\mathbb{N} = \{0, 1, 2, \dots\}$ of natural numbers to the point where they can add and multiply. Then they can solve certain equations: for example, $x + 3 = 7$ has the unique solution $x = 4$. However, they declare that, despite the efforts of their finest mathematicians, the equation $x + 3 = 2$ has no solution. We, of course, know that they have an inadequate number supply and are not aware of the existence of the negative integers. To put it another way, they have invented a system \mathbb{N} of numbers that is adequate for ordinary counting, but they must invent a larger number system \mathbb{Z} to be able to solve the equation $x + a = b$ for any a and b in \mathbb{N} .

Of course, \mathbb{Z} is also inadequate. For example, $3x = 5$ has no solution in \mathbb{Z} , and the set \mathbb{Q} of rational numbers must be invented to solve equations of the form $ax = b$. Again, the equation $x^2 = 2$ has no solution in \mathbb{Q} and so the (much) larger set \mathbb{R} of real numbers is needed. Even \mathbb{R} is deficient: $x^2 = -1$ has no solution in \mathbb{R} , which leads to the invention of the set \mathbb{C} of complex numbers. This step is in a sense the end of the story because, thanks to Gauss, we know that $f(x) = 0$ has a solution in \mathbb{C} for every polynomial f with coefficients in \mathbb{C} .

Although these number systems did not quite evolve in this way historically, the pattern is clear. When faced with an algebraic equation with no solution in a known number system, the idea is to invent a larger number system that contains

a solution. This process of adjoining solutions plays a major role in field theory. Let F be any field and let f be a polynomial in $F[x]$ that has no root in F . Then a larger field E containing F can be constructed that contains a root of f . Moreover, by repeating the process, we can find a field K containing F such that f factors completely as a product of linear factors in $K[x]$. Finally, the smallest such field (in a suitable sense) is uniquely determined by F and f and is called the splitting field of f over F . We carry out this construction in this chapter and use it, among other things, to completely classify all finite fields.

6.1 VECTOR SPACES

Order and simplification are the first steps to the mastery of a subject.

—Thomas Mann

Consider the following system of linear equations:

$$\left. \begin{array}{l} ax + by = 0 \\ cx + dy = 0 \end{array} \right\} \quad a, b, c, d \text{ real.}$$

Because the system is homogeneous (both constants on the right are zero), the set of solutions to the system has an algebraic structure. More precisely, if both

$$\begin{bmatrix} x \\ y \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

are solutions, then for any real number k , the sum and scalar product

$$\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x + x_1 \\ y + y_1 \end{bmatrix} \quad \text{and} \quad k \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} kx \\ ky \end{bmatrix}$$

are also solutions. In fact, the set of all solutions is an additive abelian group, and any such group with an appropriate scalar multiplication defined on it is an example of a real vector space. These vector spaces are the chief objects of study in linear algebra—the sister subject of abstract algebra.

In linear algebra, matrices and vector spaces over the real numbers are defined, and concepts such as basis and dimension are introduced. Most of this theory can be developed in the same way with an arbitrary field F replacing \mathbb{R} throughout, and some of this is needed in this chapter. However, a course in linear algebra is not a prerequisite to the present discussion, and this section develops just enough of the theory for the applications to fields that follow. If the reader is familiar with real vector spaces, a glance at this section will probably suffice before proceeding to Section 6.2.

Vector Spaces

If F is any field, a **vector space** V over F is an additive abelian group such that for all $a \in F$ and v in V , an element av in V is defined (called the **scalar multiple** of v by a) that satisfies the following conditions for all $a, b \in F$ and all $v, w \in V$:

$$\text{V1 } a(v + w) = av + aw.$$

$$\text{V2 } (a + b)v = av + bv.$$

$$\text{V3 } a(bv) = (ab)v.$$

$$\text{V4 } 1v = v.$$

The elements of V and F are called **vectors** and **scalars**, respectively. To emphasize the field of scalars, we call V an F -space and denote it $V = {}_F V$.

Of course, we adopt all the conventions about an additive abelian group V : The unity is called the **zero** of V , denoted 0 , and the inverse of a vector v is denoted $-v$ and is called the **negative** of v .

Example 1. If F is a field, $F^n = \{(a_1, \dots, a_n) \mid a_i \in F\}$ is a vector space with the usual componentwise addition and scalar multiplication:

$$(a_1, \dots, a_n) + (b_1, \dots, b_n) = (a_1 + b_1, \dots, a_n + b_n), \\ k(a_1, \dots, a_n) = (ka_1, \dots, ka_n).$$

If $n = 1$, $F^1 = {}_F F$ is a vector space over itself. When it is more convenient, we write the n -tuples in F^n as columns rather than rows. \square

Example 2. If F is a field, the set $M_n(F)$ of all $n \times n$ matrices over F is a vector space with the usual matrix addition, and scalar multiplication $a[a_{ij}] [= [aa_{ij}]]$ for all $a \in F$ and $[a_{ij}] \in M_n(F)$.

Example 3. Let R be any ring that contains a field F as a subring. Then $R = {}_F R$ is a vector space using the addition and multiplication of R . Thus, \mathbb{C} is an \mathbb{R} -space, and we refer repeatedly to the case where R itself is a field. Also, $F[x]$ is an F -space where F is identified with the subring of constant polynomials.

Example 4. If F is any field, the additive group $\{0\}$ is a vector space over F if we define $a \cdot 0 = 0$ for all $a \in F$. It is called the **zero space** and denoted 0 .

Theorem 1 collects several frequently used facts about vector spaces V over a field F . When no confusion can result (which is nearly always), we use the symbol 0 for both the zero of the field F and that of the additive abelian group V .

Theorem 1. Let V be an F -space where F is a field and let $a \in F$ and $v \in V$.

- (1) $0v = 0$ and $a0 = 0$.
- (2) $av = 0$ if and only if $a = 0$ in F or $v = 0$ in V .
- (3) $(-1)v = -v$.
- (4) $(-a)v = -(av) = a(-v)$.

Proof. $0v = (0+0)v = 0v + 0v$ by axiom V2, so $0v = 0$. Similarly, V1 gives $a0 = 0$, proving (1). If $av = 0$ and $a \neq 0$, then $v = 1v = (a^{-1}a)v = a^{-1}(av) = a^{-1}0 = 0$ by (1). With (1), this proves (2). As to (3), $(-1)v + v = (-1+1)v = 0v = 0$ by (1), so $(-1)v$ is the additive inverse of v . This gives (3); (4) is left as Exercise 5. \blacksquare

A subset U of a vector space ${}_F V$ is called a **subspace** of V if U is itself a vector space using the addition and scalar multiplication of V ; in other words, U is a subgroup of V that is closed under scalar multiplication ($au \in U$ for all $a \in F$ and $u \in U$). Theorem 2 is the analogue of the subgroup test. (The proof is Exercise 6.)

Theorem 2. Subspace Test. A nonempty subset U of a vector space ${}_F V$ is a subspace if and only if it is closed under addition and scalar multiplication.

Example 5. If FV is a vector space, V and $\{0\}$ are subspaces of V .

Example 6. If FV is a vector space and $v \in V$, write $Fv = \{av \mid a \in F\}$. This is easily verified to be a subspace of V (using Theorem 1).

Example 7. If A is any matrix in $M_n(F)$, show that $U = \{u \in F^n \mid Au = 0\}$ is a subspace of F^n . (Here, vectors in F^n are written as columns.)

Solution. If $u, v \in U$, then $A(u + v) = Au + Av = 0 + 0 = 0$, so $u + v \in U$ and $A(ku) = k(Au) = k0 = 0$, $k \in F$, so $ku \in U$. Hence, the subspace test applies. \square

Spanning and Independence

The most important way to describe subspaces of a vector space FV is to use the following notion: If v_1, \dots, v_n are vectors in V , a vector of the form

$$a_1v_1 + \cdots + a_nv_n \quad \text{where } a_i \in F \text{ for all } i$$

is called a **linear combination** of the v_i . The set of all such vectors is denoted

$$\text{span}\{v_1, \dots, v_n\} = \{a_1v_1 + \cdots + a_nv_n \mid a_i \in F\}.$$

We can easily verify that this is a subspace of V that contains each of the vectors v_1, v_2, \dots, v_n . Moreover, $\text{span}\{v_1, \dots, v_n\}$ is the *smallest* subspace of V containing each v_i in the sense that if U is any such subspace, then $\text{span}\{v_1, \dots, v_n\} \subseteq U$.

We use the following terminology. Let v_1, \dots, v_n be vectors in a vector space FV . Then $\text{span}\{v_1, \dots, v_n\}$ is called the subspace of V **spanned** (or **generated**) by these vectors. We say that V is **finite dimensional** if

$$V = \text{span}\{v_1, \dots, v_n\} \quad \text{for finitely many vectors } v_1, \dots, v_n.$$

In this case, we say that the vectors v_1, \dots, v_n are a **spanning set** for V .

Example 8. If F is a field, show that the space $F[x]$ is not finite dimensional.

Solution. The degree of any nonzero polynomial in $\text{span}\{f_1, \dots, f_n\}$ cannot exceed the maximum of the degrees of the (nonzero) f_i . Since $F[x]$ contains polynomials of arbitrarily large degree, it is impossible that $F[x] = \text{span}\{f_1, \dots, f_n\}$. \square

If $V = \text{span}\{v_1, \dots, v_n\}$, then every vector v in V can be written in at least one way as a linear combination of the vectors v_1, v_2, \dots, v_n . The spanning sets for which this happens in *exactly* one way for every v in V are of fundamental importance. In particular, the **trivial linear combination**

$$0v_1 + 0v_2 + \cdots + 0v_n = 0$$

is certainly one way to express the zero vector as a linear combination of the v_i , and it turns out to be enough to insist that this is the *only* way to do it.

With this in mind, a set $\{v_1, v_2, \dots, v_n\}$ of vectors in a vector space FV is called **linearly independent** (or simply **independent**) if

$$a_1v_1 + a_2v_2 + \cdots + a_nv_n = 0, \quad a_i \in F \quad \text{implies that} \quad a_1 = a_2 = \cdots = a_n = 0.$$

A set of vectors that is not independent is called **dependent**.

Example 9. If $2 \neq 0$ in the field F , show that $\{(1, 1), (1, -1)\}$ is independent in F^2 , whereas $\{(1, 2), (1, -1), (0, 1)\}$ is dependent.

Solution. If $a(1, 1) + b(1, -1) = (0, 0)$, equating first and second components gives $a + b = 0$ and $a - b = 0$. As $2 \neq 0$ and F is a field, the only solution is $a = b = 0$, so the linear combination is the trivial one. However, $-(1, 2) + (1, -1) + 3(0, 1) = (0, 0)$ shows that $\{(1, 2), (1, -1), (0, 1)\}$ is dependent. \square

The zero vector cannot belong to any independent set (by Theorem 1). On the other hand, Theorem 1 gives

Example 10. Given $v \in V$, $\{v\}$ is independent if and only if $v \neq 0$.

A set of vectors in a vector space V is called a **basis** for V if it is linearly independent and also spans V .

Example 11. In the vector space F^n , consider the vectors

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, \dots, 0), \quad \dots, \quad e_n = (0, 0, \dots, 1)$$

in F^n . We have $(a_1, a_2, \dots, a_n) = a_1e_1 + a_2e_2 + \dots + a_ne_n$ for all $a_i \in F$. It follows that $\{e_1, e_2, \dots, e_n\}$ is a basis of F^n , which is called the **standard basis**.

Theorem 3. If $\{v_1, \dots, v_n\}$ is a basis of FV , then every vector v in V has a unique representation as a linear combination $v = a_1v_1 + \dots + a_nv_n$, $a_i \in F$.

Proof. Such a representation exists because $V = \text{span}\{v_1, \dots, v_n\}$. If we have two expressions $v = a_1v_1 + \dots + a_nv_n$ and $v = b_1v_1 + \dots + b_nv_n$ for v , then

$$0 = v - v = (a_1 - b_1)v_1 + \dots + (a_n - b_n)v_n.$$

Hence, the independence of $\{v_1, \dots, v_n\}$ guarantees that $a_i = b_i$ for each i . \blacksquare

If F is a finite field with $|F| = q$, Theorem 3 shows that a vector space FV with a basis $\{v_1, \dots, v_n\}$ of n vectors has exactly q^n elements. In fact, in forming a typical vector $v = a_1v_1 + \dots + a_nv_n$ in V , there are q choices for each coefficient a_i , and Theorem 3 guarantees that each series of choices produces a different vector in V . We make use of this fact in Section 6.4 on finite fields.

Note that if V has another basis $\{w_1, \dots, w_m\}$, then, similarly, $|V| = q^m$. Hence, $q^n = |V| = q^m$ and so $n = m$. In other words, the number of elements in any two bases of V is the same. In fact, this remains true even for arbitrary fields and leads to the fundamental concept of dimension. We now turn to a proof of this basic fact.

Dimension

The concept of basis is fundamental to the theory of vector spaces, and we develop the most important properties of bases in this section. The key result is

Theorem 4. Fundamental Theorem. Suppose that $V = \text{span}\{v_1, \dots, v_n\}$ is a vector space and that $\{u_1, \dots, u_m\}$ is an independent subset of V . Then $m \leq n$.

Proof. We assume that $m > n$ and show that this leads to a contradiction. Because $V = \text{span}\{v_1, \dots, v_n\}$, write $u_1 = a_1v_1 + \dots + a_nv_n$. As $u_1 \neq 0$, not all the a_i are zero, say $a_1 \neq 0$ (after relabeling the v_i). Then $V = \text{span}\{u_1, v_2, \dots, v_n\}$ by Exercise 21.

Now write $u_2 = b_1 u_1 + a_2 v_2 + \dots + a_n v_n$. Then some a_i is nonzero because $\{u_1, u_2\}$ is independent (by Exercise 22), so $V = \text{span}\{u_1, u_2, v_3, \dots, v_n\}$ as before. As $m > n$, this procedure continues until all the vectors v_1, \dots, v_n are replaced by u_1, \dots, u_n . In particular, $V = \text{span}\{u_1, \dots, u_n\}$. But then u_{n+1} is a linear combination of u_1, \dots, u_n , a contradiction because $\{u_1, \dots, u_m\}$ is independent. ■

If $V = \text{span}\{v_1, \dots, v_n\}$, and if $\{u_1, \dots, u_m\}$ is independent in V , the proof of Theorem 4 shows that not only $m \leq n$ but also m of the (spanning) vectors v_1, \dots, v_n can be replaced by the (independent) vectors u_1, \dots, u_m and the resulting set will still span V . This result is called the **Steinitz exchange lemma**.

The first consequence of Theorem 4 is that the number of vectors in a basis of a vector space V is an **invariant** of V ; that is, it is the same for any basis.

Theorem 5. Invariance Theorem. If $\{u_1, \dots, u_m\}$ and $\{v_1, \dots, v_n\}$ are two bases of a vector space V , then $m = n$.

Proof. We have $m \leq n$ by Theorem 4 because $\{u_1, \dots, u_m\}$ is independent and $V = \text{span}\{v_1, \dots, v_n\}$. Interchanging the u_i and v_j gives $n \leq m$, so $m = n$. ■

Hence, if a vector space $V \neq 0$ has a basis $\{v_1, \dots, v_n\}$, the integer n does not depend on the choice of basis, and n is called the **dimension** of V and is denoted

$$n = \dim V.$$

The dimension of the zero space is defined to be 0. This is equivalent to regarding the zero space as having an empty basis and is consistent with the fact that the zero vector cannot belong to any independent set. Hence, the statement that $\dim V = n$ if and only if V has a basis of n vectors holds even if $n = 0$.

Example 12. $\dim_{\mathbb{R}} \mathbb{C} = 2$ because $\{1, i\}$ is a basis.

Example 13. $\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$ is a basis of $M_2(F)$, so $\dim_F M_2(F) = 4$. Similarly, $\dim_F M_n(F) = n^2$.

Example 14. If $n \geq 1$, then $\dim F^n = n$ by Example 11.

Example 15. Consider the subspace $V = \text{span}\{1, x, x^2, \dots, x^n\}$ of $F[x]$. Then $\dim V = n+1$ because $\{1, x, x^2, \dots, x^n\}$ is independent by the definition of the indeterminate x in Section 4.1. Hence $\dim(F[x])$ is not finite by Theorem 4.

The second consequence of the fundamental theorem is that any finite dimensional vector space *has* a basis. We need Lemmas 1 and 2 (Exercises 24 and 25).

Lemma 1. Let $\{v_1, \dots, v_n\}$ be an independent set in a vector space V . If $v \in V$, then $\{v, v_1, \dots, v_n\}$ is independent if and only if $v \notin \text{span}\{v_1, \dots, v_n\}$.

Lemma 2. A set of vectors is dependent if and only if one of them is in the span of the rest.

Theorem 6. Let $V \neq 0$ be a finite dimensional vector space, say V is spanned by n vectors.

- (1) V has a finite basis and $\dim V \leq n$.
- (2) Each independent subset of V is part of a basis.
- (3) Each finite spanning set for V contains a basis.

Proof. (1) Because V has a finite spanning set by hypothesis, it has a finite basis by (3), which is proved below. Because the basis is independent, $\dim V \leq n$ by Theorem 4.

(2) Let $\{v_1, \dots, v_k\}$ be independent in V , a finite set by the fundamental theorem. If $\text{span}\{v_1, \dots, v_k\} = V$, the proof is complete. Otherwise, there exists $v_{k+1} \in V$, with $v_{k+1} \notin \text{span}\{v_1, \dots, v_k\}$. Then $\{v_1, \dots, v_k, v_{k+1}\}$ is independent by Lemma 1, which completes the proof if $V = \text{span}\{v_1, \dots, v_k, v_{k+1}\}$. If not, repeat the process. Thus, either the proof is complete at some stage or the process constructs arbitrarily large independent subsets of V . But this is impossible by the fundamental theorem, because V is spanned by n vectors.

(3) Let $V = \text{span}\{v_1, \dots, v_n\}$. If $\{v_1, \dots, v_n\}$ is independent, there is nothing to prove. Otherwise, one of these vectors lies in the span of the rest by Lemma 2; relabeling if necessary, let $v_1 \in \text{span}\{v_2, \dots, v_n\}$. Then $V = \text{span}\{v_2, \dots, v_n\}$, so the proof is complete if $\{v_2, \dots, v_n\}$ is independent. If not, repeat the process. If a basis is encountered at some stage, the proof is complete. Otherwise, we ultimately reach $V = \text{span}\{v_n\}$. But then $\{v_n\}$ is a basis because $v_n \neq 0$ ($V \neq 0$ by hypothesis). ■

Parts (2) and (3) of Theorem 6 reveal a useful property of a vector space V : If $\dim V = n$, a set B of exactly n vectors in V is independent if and only if it spans V (Theorem 7). The advantage of this is that it eliminates the need to verify one or the other of these properties when we are checking that B is a basis of V .

Theorem 7. *Let V be a vector space with $\dim V = n$ and let $B \subseteq V$ be a set of exactly n vectors. If B is independent or spans V , then B is a basis of V .*

Proof. If B is independent and does not span V , then B is part of a basis of more than n vectors by Theorem 6, which contradicts Theorem 5. Similarly, if B spans V and is not independent, then B contains a basis of fewer than n vectors by Theorem 6, again contrary to Theorem 5. ■

Example 16. Let a be an element of a field F and let $n \geq 0$. Given f in $F[x]$, with $\deg f \leq n$, show that a_0, a_1, \dots, a_n exist in F such that

$$f = a_0 + a_1(x - a) + \cdots + a_n(x - a)^n.$$

Solution. Let $V = \text{span}\{1, x, \dots, x^n\}$ in $F[x]$. If $B = \{1, (x - a), \dots, (x - a)^n\}$, then $B \subseteq V$ and we show that B spans V . As $\dim V = n + 1$, it suffices by Theorem 7 to show that B is independent. Suppose that $r_0 + r_1(x - a) + \cdots + r_n(x - a)^n = 0$ in $F[x]$, $r_i \in F$. Then r_n is the coefficient of x^n on the left-hand side, so $r_n = 0$. Next $r_{n-1} = 0$ in the same way, and we continue to get $r_i = 0$ for all i . □

We conclude this section with a theorem relating the dimension of a vector space V to the dimensions of its subspaces.

Theorem 8. *Let $V \neq 0$ be a vector space, $\dim V = n$, and let $U \subseteq V$ be a subspace.*

- (1) *U has a basis and $\dim U \leq n$.*
- (2) *If $\dim U = n$, then $U = V$.*
- (3) *Every basis for U is part of a basis for V .*

Proof. (1) If $U = 0$, then it has an empty basis and $\dim U = 0$. If $U \neq 0$, let $u_1 \neq 0$ in U . If $\text{span}\{u_1\} = U$ then $\{u_1\}$ is a basis of U . Otherwise, the construction

in the proof of Theorem 6(2) either produces a basis for U or creates arbitrarily large independent subsets of V . The latter outcome cannot happen by Theorem 4 because V is spanned by n vectors. Hence, U has a basis $\{u_1, \dots, u_m\}$. Then $\dim U = m$, and $m \leq n$ again by Theorem 4.

(2) If $\dim U = n$, any basis $\{u_1, \dots, u_n\}$ of U is a basis of V by Theorem 7. Thus, $U = \text{span}\{u_1, \dots, u_n\} = V$.

(3) This follows from Theorem 6. ■

Exercises 6.1

Throughout these exercises, F denotes a field.

1. Which of the following are subspaces of F^3 ? Support your answer.
 - (a) $U = \{(a, b, 1) \mid a, b \in F\}$
 - (b) $U = \{(a, b, c) \mid a - 2b + 3c = 0\}$
 - (c) $U = \{(a, b - 1, c) \mid a, b, c \in F\}$
 - (d) $U = \{(2a + b, b - c, 3b + a) \mid a, b, c \in F\}$
2. Which of the following are subspaces of $F[x]$? Support your answer.
 - (a) $U = \{f \mid f(2) = 0\}$
 - (b) $U = \{xf \mid f \in F[x]\}$
 - (c) $U = \{f \mid \deg f \leq 3\}$
 - (d) $U = \{f \mid f(3) = 1\}$
3. Show that $F^3 = \text{span}\{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}$ provided that $2 \neq 0$ in F .
4. (a) Show that $\text{span}\{u, v, w\} = \text{span}\{u + v, u + w, v + w\}$ in any vector space $_F V$ where $2 \neq 0$ in F .
 - (b) Is (a) true if $F = \mathbb{Z}_2$? Support your answer.
5. Prove (4) of Theorem 1.
6. Prove the subspace test (Theorem 2).
7. Which of the following are independent in V ? Support your answer.
 - (a) $\{(1, 2, 3), (4, 0, 1), (2, 1, 0)\}$ in $V = \mathbb{Z}_5^3$
 - (b) $\{(1, 0, 1, 0), (1, 1, 0, 0), (0, 1, 0, 1), (0, 0, 1, 1)\}$ in $V = F^4$
 - (c) $\{x^2 + 1, x + 1, x\}$ in $V = F[x]$
 - (d) $\{x^2 - x + 1, 2x^2 + x + 1, x - 1\}$ in $V = F[x]$
8. Given $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in $M_2(F)$, show that A is invertible if and only if $\{(a, b), (c, d)\}$ is a basis of F^2 .
9. (a) Show that $\{1, \sqrt{2}, \sqrt{3}\}$ is independent in \mathbb{R} over \mathbb{Q} .
 - (b) Show that $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$ is independent in \mathbb{R} over \mathbb{Q} .

[Hint: $(c + d\sqrt{2})(c - d\sqrt{2}) = c^2 - 2d^2$.]
10. Find a basis of \mathbb{R}^2 containing $v = (1, -2)$ and a basis not containing v .
11. Find infinitely many bases of \mathbb{R}^3 containing $v = (1, -1, 0)$ and $w = (1, 1, 1)$.
12. Find all values of r for which $\{(2, r, 1), (1, 0, 2), (0, 1, -2)\}$ is independent in \mathbb{R}^3 .
13. Suppose that f and g in $F[x]$ satisfy $f(a) = 0 = g(b)$, $f(b) \neq 0$, and $g(a) \neq 0$ for some fixed a and b in F . Show that $\{f, g\}$ is independent in $F[x]$.
14. Show that $\{f_1, f_2, \dots, f_n\}$ is independent in $F[x]$ whenever $\deg f_1, \deg f_2, \dots, \deg f_n$ are distinct. [See solution to Example 16.]
15. If A is a 2×2 matrix in $M_2(F)$, show that $a_0I + a_1A + a_2A^2 + a_3A^3 + a_4A^4 = 0$ for some $a_i \in F$, not all zero.

16. If $\{A_1, A_2, \dots, A_k\}$ is linearly independent in $M_n(F)$, and if U is invertible, show that $\text{span}\{A_1U, A_2U, \dots, A_kU\}$ has dimension k .
17. Let $\{A_1, \dots, A_k\}$ in $M_n(F)$ be such that, for some column $v \neq 0$ in F^n , $A_i v = 0$ for each i . Show that $\text{span}\{A_1, \dots, A_k\} \neq M_n(F)$.
18. If $X = \{1, 2, \dots, n\}$, let $D_n = \{f \mid f: X \rightarrow F \text{ is a mapping}\}$. If $f, g \in D_n$ and $a \in F$, define the pointwise sum $f + g$ and scalar product af by $(f + g)(x) = f(x) + g(x)$ and $(af)(x) = a f(x)$ for all $x \in X$. Show that D_n is a vector space over F and that $\dim D_n = n$.
19. Let R be a ring such that the field $F \subseteq R$ is a subring. Assume that $\dim(FR) = n$.
- If $r \in R$, show that $p(r) = 0$ for some polynomial $p \neq 0$ in $F[x]$, with $\deg p \leq n$. [Hint: Can $\{1, r, r^2, \dots, r^n\}$ be independent?]
 - If R is an integral domain, show that it must be a field.
20. Let $F \subseteq E$ be fields with $\dim_F E = n$. If $_E V$ is a vector space over E (and hence over F), and if $\dim_E V = m$, show that $\dim_F V = mn$.
21. If $u = a_1v_1 + a_2v_2 + \dots + a_nv_n$ in a vector space V , and if $a_1 \neq 0$, show that $\text{span}\{v_1, v_2, \dots, v_n\} = \text{span}\{u, v_2, \dots, v_n\}$.
22. (a) Show that every subset of an independent set of vectors is independent.
(b) Show that a set of vectors is dependent if it contains a dependent subset.
23. (a) Show that an independent set $\{v_1, \dots, v_n\}$ in $_F V$ with n maximal is a basis.
(b) Show that a spanning set $\{v_1, \dots, v_n\}$ of $_F V$ with n minimal is a basis.
24. Prove Lemma 1.
25. Prove Lemma 2.
26. Let U and W be subspaces of a finite dimensional vector space V .
- Show that $U + W = \{u + w \mid u \in U, w \in W\}$ is a subspace of V .
 - If $U \cap W = 0$, show that $\dim(U + W) = \dim U + \dim W$.
 - In general, show that $\dim(U + W) = \dim U + \dim W - \dim(U \cap W)$.
[Hint: By Theorem 6(2), extend a basis of $U \cap V$ to bases of U and of W .]
27. If U and W are finite dimensional subspaces of V , show that $U + W$ (defined in Exercise 26) is finite dimensional.
28. A polynomial p in $F[x]$ is called **even** if $p(-x) = p(x)$, and p is **odd** if $p(-x) = -p(x)$. Let $V = \text{span}\{1, x, x^2, \dots, x^n\}$ and let U and W denote, respectively, the sets of even and odd polynomials in V . Assume that $2 \neq 0$ in F .
- Show that U and W are subspaces of V such that $U \cap W = 0$ and $U + W = V$ (see Exercise 26).
 - Find $\dim U$ and $\dim W$.
29. Let U be a subspace of a vector space V with $\dim U = m$ and let $v \in V$. Given $W = \{u + av \mid u \in U, a \in F\}$, show that W is a subspace of V and that $\dim W = m$ or $\dim W = m + 1$.
30. If U is a subspace of a vector space $_F V$, define a scalar multiplication on the (additive) factor group V/U by $a(v + U) = av + U$. Show that V/U is a vector space and that if V is finite dimensional, then V/U is finite dimensional and $\dim V/U = \dim V - \dim U$.
31. A **linear transformation** $\varphi: {}_F V \rightarrow {}_F W$ is a map such that $\varphi(v + w) = \varphi(v) + \varphi(w)$ and $\varphi(av) = a\varphi(v)$ for all $a \in F$ and all $v, w \in V$.
- Show that $\ker \varphi$ and $\text{im } \varphi$ are subspaces of V and W , respectively.
 - If V is finite dimensional, show that $\text{im } \varphi$ is also finite dimensional.
 - If V is finite dimensional, show that $\dim V = \dim(\ker \varphi) + \dim(\text{im } \varphi)$. (This is called the **dimension theorem**.) [Hint: Extend a basis $\{u_1, \dots, u_m\}$ of $\ker \varphi$ to a basis $\{u_1, \dots, u_m, v_1, \dots, v_k\}$ of V .]

- 32.** Vector spaces FV and FW are called isomorphic (written $V \cong W$) if a one-to-one, onto linear transformation $V \rightarrow W$ exists (see Exercise 31). If FV has dimension n , show that $V \cong F^n$.

6.2 ALGEBRAIC EXTENSIONS

Much of field theory concerns the relationship between two fields F and E , with $E \supseteq F$. Of course, this is taken to mean that F is a subring of E , and in this case, F is called a **subfield** of E and E is called an **extension field** of F (or simply an **extension** of F). Everything we do relies on the fact that $E = FE$ is a vector space over F using the operations of E . If the vector space FE has finite dimension, then E is called a **finite extension** of F , and we write $\dim FE = [E : F]$.

Example 1. $\mathbb{C} \supseteq \mathbb{R}$ is finite, and $[\mathbb{C} : \mathbb{R}] = 2$ because $\{1, i\}$ is an \mathbb{R} -basis of \mathbb{C} .

Example 2. We demonstrate later that $\mathbb{R} \supseteq \mathbb{Q}$ is not a finite extension.

Theorem 1. Let $E \supseteq F$ be a finite extension with $[E : F] = n$. If $u \in E$, a polynomial $f \neq 0$ in $F[x]$ exists such that $\deg f \leq n$ and $f(u) = 0$.

Proof. The $n + 1$ elements $1, u, u^2, \dots, u^n$ of E cannot be F -independent because $\dim FE = n$ (Theorem 4 §6.1). Hence, $a_0 + a_1u + a_2u^2 + \dots + a_nu^n = 0$ for some $a_i \in F$, not all zero. Take $f = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$. ■

If $E \supseteq F$ is an extension of fields, an element $u \in E$ is said to be **algebraic** over F if $f(u) = 0$ for some polynomial $f \neq 0$ in $F[x]$ (which may be taken to be monic since F is a field). An extension $E \supseteq F$ is called an **algebraic extension** if every element of E is algebraic over F . Thus, Theorem 1 asserts that every finite extension is algebraic. We show later (Example 16) that the converse is not true.

Example 3. The complex numbers $\sqrt[3]{2}$ and i are algebraic over \mathbb{Q} because they are roots of $x^3 - 2$ and $x^2 + 1$, respectively.

Example 4. Each element a of F is algebraic over F , being a root of $x - a$.

Example 5. The number $u = \sqrt{2} - \sqrt{3}$ is algebraic over \mathbb{Q} . Indeed $u^2 = 5 - 2\sqrt{6}$, so $(u^2 - 5)^2 = 24$. This gives $u^4 - 10u^2 + 1 = 0$, so u is a root of $x^4 - 10x^2 + 1$.

If $E \supseteq F$ are fields, an element $u \in E$ that is *not* algebraic over F is called **transcendental** over F . The reader should not get the idea that all complex numbers are algebraic over \mathbb{Q} . This is far from the case, although establishing that a given number is transcendental is usually difficult. The next theorem, which we state without proof,⁷¹ identifies two transcendental numbers.

Theorem 2. The numbers π and e (from calculus) are transcendental.

In 1873, Charles Hermite gave the first proof that e is transcendental over \mathbb{Q} . This proof stimulated interest in such questions and, in 1882, Ferdinand Lindemann succeeded in proving that π is also transcendental over \mathbb{Q} . This result is famous,

⁷¹For proofs of these assertions and more information on transcendental numbers, see Niven, I., *Irrational Numbers*, Carus Monograph II, Washington, DC: Mathematical Association of America, 1956.

partly because it settled a classical question dating back to the ancient Greeks: Is it possible, using only compass and straightedge, to square the circle (construct a square whose area equals that of a given circle)? The answer is no, because the existence of such a construction implies that π is algebraic (see Section 6.5).

These results are difficult, so it is surprising that there are “more” transcendental complex numbers (over \mathbb{Q}) than there are algebraic ones. In fact, it can be shown that the set \mathbb{A} of complex numbers that are algebraic over \mathbb{Q} is “countable” in the sense that there exists a bijection $\mathbb{A} \rightarrow \mathbb{N}$. However, \mathbb{C} is “uncountable” in the sense that no one-to-one map $\mathbb{C} \rightarrow \mathbb{N}$ exists. This was first established in 1874 by Georg Cantor, and he gave another proof in 1891 using his celebrated “diagonal method.”

If $E \supseteq F$ are fields, and if u_1, \dots, u_n are elements of E , let $F(u_1, \dots, u_n)$ denote the intersection of all subfields of E that contain F and also contain each of the elements u_i . More formally,

$$F(u_1, \dots, u_n) = \cap \{K \supseteq F \mid K \text{ is a subfield of } E \text{ containing each } u_i\}.$$

It is easy to verify that this is again a field containing F and all the u_i . Thus, it is the *smallest* such subfield of E (in the sense that it is contained in every such subfield). The field $F(u_1, \dots, u_n)$ is called the subfield of E generated over F by the elements u_1, \dots, u_n . If $u \in E$, the extension $F(u)$ is called a **simple extension** of F in E .

Example 6. Show that $\mathbb{R}(i) = \mathbb{C}$.

Solution. \mathbb{C} is certainly a field containing \mathbb{R} and i , so $\mathbb{R}(i) \subseteq \mathbb{C}$. However, given $z = a + bi$ in \mathbb{C} , where $a, b \in \mathbb{R}$, z lies in any field containing \mathbb{R} and i and, in particular, $z \in \mathbb{R}(i)$. Thus, $\mathbb{C} \subseteq \mathbb{R}(i)$, so $\mathbb{C} = \mathbb{R}(i)$ as asserted. \square

Example 7. Show that $\mathbb{Q}(i, -i) = \mathbb{Q}(i)$.

Solution. $\mathbb{Q}(i) \subseteq \mathbb{Q}(i, -i)$ because $\mathbb{Q}(i, -i)$ contains \mathbb{Q} and i . But $\mathbb{Q}(i)$ is a field containing \mathbb{Q} and i , and so it also contains $-i$. Hence, $\mathbb{Q}(i, -i) \subseteq \mathbb{Q}(i)$. \square

Example 8. Show that $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$.

Solution. Write $E = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$. Clearly, E is contained in any subfield containing \mathbb{Q} and $\sqrt{2}$; in particular, $E \subseteq \mathbb{Q}(\sqrt{2})$. On the other hand, E is a field (by Example 4 §3.2), $\mathbb{Q} \subseteq E$ and $\sqrt{2} \in E$, so $\mathbb{Q}(\sqrt{2}) \subseteq E$. Hence, $E = \mathbb{Q}(\sqrt{2})$. \square

Example 9. If $E \supseteq F$ are fields and $u \in E$, then $F(u) = F$ if and only if $u \in F$.

If $E \supseteq F$ are fields and u and v are elements of E , then $E \supseteq F(u)$ is also an extension of fields. Thus, we can speak of $F(u)(v)$. This is evidently a subfield of E containing F, u , and v , and so $F(u, v) \subseteq F(u)(v)$. In fact, this is equality: The subfield $F(u, v)$ contains $F(u)$ (because it contains F and u), and it also contains v , so $F(u)(v) \subseteq F(u, v)$. A similar argument proves the next result (Exercise 25).

Lemma 1. Let $E \supseteq F$ be fields and let u_1, u_2, \dots, u_n be elements of E , $n \geq 2$. Then for $1 \leq k \leq n - 1$, we have

$$F(u_1, \dots, u_k)(u_{k+1}, \dots, u_n) = F(u_1, \dots, u_k, \dots, u_n).$$

For fields $E \supseteq F$, Lemma 1 implies that the subfield $F(u_1, \dots, u_n)$ generated over F by u_1, \dots, u_n can be built up as a chain of simple extensions:

$$\begin{aligned} F(u_1, u_2) &= F(u_1)(u_2), \\ F(u_1, u_2, u_3) &= F(u_1, u_2)(u_3), \\ &\vdots && \vdots \\ F(u_1, \dots, u_{n-1}, u_n) &= F(u_1, \dots, u_{n-1})(u_n). \end{aligned}$$

This highlights the importance of studying simple extensions $F(u)$.

If u is transcendental over F , it is routine to verify (Exercise 31) that

$$F(u) = \{f(u)g(u)^{-1} \mid f, g \text{ in } F[x]; g \neq 0\}.$$

Hence, $F(u) \cong F(x)$ —the field of quotients of the integral domain $F[x]$. However, our interest lies in simple extensions $F(u)$, where u is algebraic over F .

Theorem 1 asserts that every element u of a finite extension E of F is algebraic over F . This has a partial converse: If $E \supseteq F$ is a field extension, and if $u \in E$ is algebraic over F , then u belongs to a finite extension of F contained in E . Indeed, we show that $F(u)$ is a finite extension of F containing u . We present this fact, along with an explicit description of $F(u)$, in Theorem 4. However, the proof involves another important notion.

Let u be algebraic over F , where u lies in some extension of F . Then $f(u) = 0$ for some nonzero polynomial $f \in F[x]$, which we may assume to be monic.

Theorem 3. *Let u be algebraic over F . Choose a monic polynomial m of minimal degree such that $m(u) = 0$. Then*

- (1) *m is irreducible in $F[x]$.*
- (2) *If f is in $F[x]$, then $f(u) = 0$ if and only if $m|f$.*
- (3) *m is uniquely determined by u .*

Proof. (1) Suppose that $m = fg$ in $F[x]$, where $\deg f < \deg m$ and $\deg g < \deg m$. Then $f(u)g(u) = m(u) = 0$ implies that $f(u) = 0$ or $g(u) = 0$, a contradiction.

(2) If $f(u) = 0$, use the division algorithm (Theorem 4 §4.1) to write $f = qm + r$ in $F[x]$, where $r = 0$ or $\deg r < \deg m$. Then $r(u) = f(u) - q(u)m(u) = 0$, so $r \neq 0$ would contradict the choice of m . Thus, $r = 0$ and $m|f$. The converse is clear.

(3) Let m' be another monic polynomial of minimal degree with $m'(u) = 0$. Then $m|m'$ by (2). Moreover, $m'|m$ because (2) is also true of m' . Thus, $m = m'$ because both are monic (Theorem 9 §4.2). \blacksquare

If u is algebraic over F , the polynomial m in Theorem 3 is called the **minimal polynomial** of u over F . Since m is uniquely determined by u and F , we are entitled to define the **degree** of u over F by $\deg_F(u) = \deg(m)$. Note that if $p \in F[x]$ is monic, irreducible and $p(u) = 0$, then $p = m$ by Theorem 3(2).

Example 10. Find the minimal polynomial of $u = \sqrt{1 + \sqrt{3}}$ over \mathbb{Q} .

Solution. We have $u^2 = 1 + \sqrt{3}$, so $(u^2 - 1)^2 = 3$, that is $u^4 - 2u^2 - 2 = 0$. The polynomial $x^4 - 2x^2 - 2$ is irreducible in $\mathbb{Q}[x]$ by the Eisenstein criterion. Hence $m = x^4 - 2x^2 - 2$ by Theorem 3, so $\deg_{\mathbb{Q}}(u) = 4$. \square

The minimal polynomial provides a lot of information about an element u algebraic over a field F . In particular it gives a useful description of the simple extension $F(u)$ generated by u over F , and will be referred to several times.

Theorem 4. If $E \supseteq F$ are fields, let $u \in E$ be algebraic over F of degree n .

- (1) $F(u) = \{a_0 + a_1u + \cdots + a_{n-1}u^{n-1} \mid a_i \in F, n \geq 0\} = \{f(u) \mid f \in F[x]\}$.
- (2) $\{1, u, \dots, u^{n-1}\}$ is an F -basis of $F(u)$, so $[F(u) : F] = n = \deg_F(u)$.
- (3) $F(u) \cong F[x]/\langle m \rangle$, where m is the minimal polynomial of u over F .

Proof. Define $\theta : F[x] \rightarrow E$ by $\theta(f) = f(u)$. Then θ is a ring homomorphism and $\ker \theta = \{f \in F[x] \mid f(u) = 0\} = \langle m \rangle$ by Theorem 3, where m is the minimal polynomial of u over F . Then, by the ring isomorphism theorem (Theorem 4 §3.4),

$$\frac{F[x]}{\langle m \rangle} \cong \text{im } \theta = \{f(u) \mid f \in F[x]\}.$$

Now $F(u)$ is a field containing F and u , and so contains $f(u)$ for all $f \in F[x]$. Hence, $\text{im } \theta \subseteq F(u)$. But, $F[x]/\langle m \rangle$ is a field because m is irreducible (Theorem 3 §4.3), so $\text{im } \theta$ is a field. Because $\text{im } \theta$ contains F and u , this shows that $F(u) \subseteq \text{im } \theta$. Thus, $F(u) = \text{im } \theta$, which proves (1) and (3).

It remains to show that $B = \{1, u, \dots, u^{n-1}\}$ is an F -basis of $F(u)$. To show that B is independent, let

$$a_0 + a_1u + \cdots + a_{n-1}u^{n-1} = 0, \quad a_i \in F.$$

Then $g(u) = 0$, where $g = a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$ so $g \neq 0$ in $F[x]$ would contradict the choice of the minimal polynomial m . Hence $g = 0$, so $a_i = 0$ for all i . Thus B is independent. To show that B spans $F(u)$, let $f(u) \in F(u)$ and write $f = qm + r$ in $F[x]$ where, since $\deg m = n$, r has the form $r = b_0 + b_1x + \cdots + b_{n-1}x^{n-1}$, $b_i \in F$. As $m(u) = 0$, we get $f(u) = r(u) = b_0 + b_1u + \cdots + b_{n-1}u^{n-1}$. Thus B spans $F(u)$ and the proof is complete. ■

Note that Theorem 4(3) shows that the field F and the minimal polynomial of the algebraic element u completely determine the extension $F(u)$, and hence that $F(u) \cong F(v)$ whenever u and v have the same minimal polynomial over F .

The description of $F(u)$ in Theorem 4(1) makes it clear how to add in $F(u)$. However, multiplying requires the minimal polynomial, as Example 11 demonstrates.

Example 11. Describe the multiplication in $\mathbb{Q}(u)$ where $u = 1 + i$.

Solution. We have $(u - 1)^2 = i^2 = -1$, so $u^2 - 2u + 2 = 0$. Write $m = x^2 - 2x + 2$. Since m is irreducible over \mathbb{Q} (it has no root in \mathbb{Q}), it is the minimal polynomial of u . Hence, $\mathbb{Q}(u) = \{a + bu \mid a, b \in \mathbb{Q}\}$ by Theorem 4, and as $u^2 = 2u - 2$,

$$\begin{aligned} (a + bu)(a' + b'u) &= aa' + (ab' + ba')u + bb'u^2 \\ &= (aa' - 2bb') + (ab' + ba' + 2bb')u. \end{aligned}$$

This describes the multiplication in $\mathbb{Q}(u)$. □

Example 12. Show that $[\mathbb{R} : \mathbb{Q}]$ is not finite.

Solution. The polynomial $x^n - 2$ is irreducible over \mathbb{Q} for any $n \geq 1$ by the Eisenstein criterion (Theorem 8 §4.2). If we write $E = \mathbb{Q}(\sqrt[n]{2})$, this means that $[E : \mathbb{Q}] = n$

by Theorem 4. Thus, $\mathbb{Q}\mathbb{R}$ contains subspaces of arbitrarily large dimension and so cannot be finite dimensional by Theorem 4 §6.1. \square

Before proceeding, we require the following basic result about finite extensions.

Theorem 5. Multiplication Theorem. *Let $K \supseteq E \supseteq F$ be fields. Then $[K : F]$ is finite if and only if both $[K : E]$ and $[E : F]$ are finite. In this case,*

$$[K : F] = [K : E] \cdot [E : F].$$

Moreover, if $\{e_1, \dots, e_m\}$ is any F -basis of ${}_F E$ and $\{k_1, \dots, k_n\}$ is any E -basis of ${}_E K$, then

$$B = \{e_i k_j \mid 1 \leq i \leq m, 1 \leq j \leq n\}$$

is an F -basis of K .

Proof. If $[K : F]$ is finite, then $[E : F]$ is also finite by Theorem 8 §6.1 because ${}_F E$ is a subspace of ${}_F K$. Next, any F -basis of ${}_F K$ is certainly an E -spanning set of ${}_E K$, so $[K : E]$ is finite by Theorem 6 §6.1.

Conversely, in the notation of the Theorem, it suffices to prove that B is an F -basis of ${}_F K$. First, B spans ${}_F K$. For if $c \in K$, then, because $\{k_1, \dots, k_n\}$ is an E -basis of K , write $c = \sum_{j=1}^n b_j k_j$, where $b_j \in E$ for each j . But then, for each j , $b_j = \sum_{i=1}^m a_{ij} e_i$, where $a_{ij} \in F$ for all i and j . Combining these expressions gives

$$c = \sum_{j=1}^n (\sum_{i=1}^m a_{ij} e_i) k_j = \sum_{j=1}^n \sum_{i=1}^m a_{ij} e_i k_j.$$

It follows that B spans ${}_F K$. Finally, to prove that B is F -independent, let

$$\sum_{i=1}^m \sum_{j=1}^n a_{ij} e_i k_j = 0$$

where $a_{ij} \in F$ for all i and j . Then $\sum_{j=1}^n (\sum_{i=1}^m a_{ij} e_i) k_j = 0$, so, as $\{k_1, \dots, k_n\}$ is E -independent, $\sum_{i=1}^m a_{ij} e_i = 0$ for each j . But then $a_{ij} = 0$ for all i and j because $\{e_1, \dots, e_m\}$ is F -independent. Hence, B is F -independent. \blacksquare

The multiplication theorem gives a divisibility relationship between dimensions, so it plays a role in field theory somewhat analogous to the role of Lagrange's theorem for groups. Consequently, we refer to it constantly, both in this chapter and in Chapter 10 on Galois theory.

Corollary. *Let $E \supseteq F$ be fields and let $u \in E$ be algebraic over F . If $v \in F(u)$, then v is also algebraic over F and $\deg_F(v)$ divides $\deg_F(u)$.*

Proof. Here, $F(u) \supseteq F(v) \supseteq F$ and $F(u) \supseteq F$ is finite, so $v \in F(u)$ is algebraic over F by Theorem 1. Also, $[F(u) : F] = [F(u) : F(v)][F(v) : F]$ by Theorem 5, so we are done because $\deg_F(u) = [F(u) : F]$ and $\deg_F(v) = [F(v) : F]$. \blacksquare

Example 13. If $u = \sqrt[3]{2}$, show that $\mathbb{Q}(u) = \mathbb{Q}(u^2)$.

Solution. We have $\mathbb{Q}(u) \supseteq \mathbb{Q}(u^2) \supseteq \mathbb{Q}$, and $[\mathbb{Q}(u) : \mathbb{Q}] = \deg_{\mathbb{Q}}(u) = 3$ because $x^3 - 2$ is irreducible in $\mathbb{Q}[x]$. Hence, $[\mathbb{Q}(u^2) : \mathbb{Q}] = 1$ or 3 by the multiplication theorem. But $[\mathbb{Q}(u^2) : \mathbb{Q}] \neq 1$ because $u^2 \notin \mathbb{Q}$, so $[\mathbb{Q}(u^2) : \mathbb{Q}] = 3$. Thus, $\mathbb{Q}(u) = \mathbb{Q}(u^2)$ by Theorem 8 §6.1. \blacksquare

Example 14. Let $E \supseteq F$ be fields and let $u, v \in E$. If u and $u + v$ are algebraic over F , show that v is algebraic over F .

Solution. Write $L = F(u+v)$ so that $L(u) = F(u, v)$. Hence, $v \in L(u)$ so it suffices (by Theorem 1) to show that $L(u) \supseteq F$ is finite. We have the chain of fields $L(u) \supseteq L \supseteq F$. But $L \supseteq F$ is finite by Theorem 4 because $u+v$ is algebraic over F , and $L(u) \supseteq L$ is finite because u is algebraic over L (being algebraic over F). Hence, $L(u) \supseteq F$ is finite by the multiplication theorem as required. \square

Example 15. Let $E = \mathbb{Q}(\sqrt{2}, \sqrt{5})$. Find $[E : \mathbb{Q}]$, exhibit a \mathbb{Q} -basis of E , and show that $E = \mathbb{Q}(\sqrt{2} + \sqrt{5})$. Then find the minimal polynomial of $\sqrt{2} + \sqrt{5}$ over \mathbb{Q} .

Solution. We write $L = \mathbb{Q}(\sqrt{2})$ for convenience so that $E = L(\sqrt{5})$ by Lemma 1. Now $x^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over \mathbb{Q} (it has no root in \mathbb{Q}), so Theorem 4 shows that $\{1, \sqrt{2}\}$ is a \mathbb{Q} -basis of L . We claim that $x^2 - 5$ is the minimal polynomial of $\sqrt{5}$ over L . Because $\sqrt{5}$ and $-\sqrt{5}$ are the only roots of $x^2 - 5$ in \mathbb{R} , we merely need to show that $\sqrt{5} \notin L$. But $\sqrt{5} \in L$ means that $\sqrt{5} = a + b\sqrt{2}$, with $a, b \in \mathbb{Q}$ (and $a \neq 0, b \neq 0$), which implies (by squaring) that $\sqrt{2} \in \mathbb{Q}$, a contradiction. Hence, $\{1, \sqrt{5}\}$ is an L -basis of $E = L(\sqrt{5})$ over L . As $E \supseteq L \supseteq \mathbb{Q}$, it also follows from the multiplication theorem that $\{1, \sqrt{2}, \sqrt{5}, \sqrt{10}\}$ is a \mathbb{Q} -basis of E , and so $[E : \mathbb{Q}] = 4$.

We now write $u = \sqrt{2} + \sqrt{5}$. Then $u^2 = 7 + 2\sqrt{10}$, so $u^3 = 17\sqrt{2} + 11\sqrt{5}$. In particular, $17\sqrt{2} + 11\sqrt{5} \in \mathbb{Q}(u)$ and $\sqrt{2} + \sqrt{5} = u \in \mathbb{Q}(u)$. Because $\mathbb{Q}(u)$ is a field, the reader can verify that both $\sqrt{2}$ and $\sqrt{5}$ are in $\mathbb{Q}(u)$, so $E \subseteq \mathbb{Q}(u)$. The reverse inclusion is obvious, so $E = \mathbb{Q}(u)$. Hence $[\mathbb{Q}(u) : \mathbb{Q}] = [E : \mathbb{Q}] = 4$, so the minimal polynomial m of u over \mathbb{Q} has degree 4. But $u^2 = 7 + 2\sqrt{10}$ yields $(u^2 - 7)^2 = 40$, from which $u^4 - 14u^2 + 9 = 0$. From Theorem 3, m divides $x^4 - 14x^2 + 9$, so $m = x^4 - 14x^2 + 9$ because both are monic of degree 4. Incidentally, this shows that $x^4 - 14x^2 + 9$ is irreducible over \mathbb{Q} . \square

Example 15 shows that $\mathbb{Q}(\sqrt{2}, \sqrt{5})$ is in fact a simple extension of \mathbb{Q} . More generally, we can show that $\mathbb{Q}(u, v)$ is a simple extension whenever u and v are algebraic over \mathbb{Q} . In fact, this result holds for any field of characteristic 0. In this generality it is called the **primitive element theorem** (see Theorem 6 §10.1).

We now turn to a characterization of the finite extensions of a field F as exactly those obtained by adjoining finitely many algebraic elements.

Theorem 6. A field extension $E \supseteq F$ is finite if and only if $E = F(u_1, \dots, u_n)$, where each $u_i \in E$ is algebraic over F .

Proof. First, assume that $[E : F]$ is finite and proceed by induction on $[E : F]$. If $[E : F] = 1$, then $E = F = F(1)$. If $[E : F] > 1$, choose $u \in E, u \notin F$. Then $[F(u) : F] > 1$ because $u \notin F$, so the multiplication theorem gives

$$[E : F(u)] = \frac{[E : F]}{[F(u) : F]} < [E : F].$$

Since $E \supseteq F(u)$ is finite, we get $E = F(u)(u_1, \dots, u_n) = F(u, u_1, \dots, u_n)$ by induction. The elements u, u_1, \dots, u_n are algebraic over F by Theorem 1.

Conversely, if $E = F(u_1, \dots, u_n)$, where the u_i are algebraic over F , we again use induction on n . If $n = 1$, then $[E : F]$ is finite by Theorem 4. If $n > 1$, write $L = F(u_1, \dots, u_{n-1})$. Then $E \supseteq L \supseteq F$ and $[L : F]$ is finite by induction. But $E = L(u_n)$, so $[E : L]$ is finite by Theorem 4 because u_n is algebraic over F , and hence over L . Thus, $[E : F]$ is finite by the multiplication theorem. \blacksquare

The first consequence of Theorem 6 is the version of the first part of the multiplication theorem for algebraic rather than finite extensions.

Corollary 1. If $K \supseteq E \supseteq F$ are fields, then $K \supseteq F$ is an algebraic extension if and only if both $K \supseteq E$ and $E \supseteq F$ are algebraic.

Proof. Assume that $K \supseteq E$ and $E \supseteq F$ are algebraic extensions. If $u \in K$, we show that u is algebraic over F by showing that u lies in some finite extension of F (and invoking Theorem 1). Because $K \supseteq E$ is algebraic, let $f(u) = 0$, where $0 \neq f \in E[x]$. If $f = e_0 + e_1x + \dots + e_nx^n$, take $L = F(e_0, \dots, e_n)$. Then $u \in L(u)$ and we claim that $L(u) \supseteq F$ is finite. As $L(u) \supseteq L \supseteq F$, this follows by the multiplication theorem because (1) $L(u) \supseteq L$ is finite (since $f \in L[x]$), and (2) $L \supseteq F$ is finite (the e_i are algebraic because $E \supseteq F$ is algebraic by hypothesis). Hence, $K \supseteq F$ is algebraic. The converse is routine and is left to the reader. ■

Corollary 2. If $E \supseteq F$ are fields, let $A = \{u \in E \mid u \text{ is algebraic over } F\}$. Then A is a subfield of E and so is an algebraic extension of F .

Proof. To prove that A is a subfield of E , it is enough to show that any two elements u and v of A lie in some finite extension $L \supseteq F$ (then L is algebraic over F by Theorem 1). But $L = F(u, v)$ fills the bill by Theorem 6. ■

The field A in Corollary 2 is called the **algebraic closure** of F in E . Clearly, it is the largest algebraic extension of F that is contained in E . The following special case will be referred to frequently. The field

$$\mathbb{A} = \{u \in \mathbb{C} \mid u \text{ is algebraic over } \mathbb{Q}\}$$

is called the **field of algebraic numbers**. The field \mathbb{A} shows that while every finite extension is algebraic (by Theorem 1), the converse is not true.

Example 16. Show that \mathbb{A} is an algebraic extension of \mathbb{Q} that is not finite.

Solution. Clearly, $\mathbb{A} \supseteq \mathbb{Q}(\sqrt[3]{2}) \supseteq \mathbb{Q}$, so the argument in Example 12 works. □

Exercises 6.2

Throughout these exercises, F denotes a field.

1. In each case, show that $u \in \mathbb{C}$ is algebraic over \mathbb{Q} .
 - (a) $u = \sqrt{3} + \sqrt{5}$
 - (b) $u = 1 + \sqrt{1 + \sqrt[3]{2}}$
 - (c) $u = \sqrt{\sqrt{3} - 2i}$
 - (d) $u = v + v^2$, where $v = \sqrt[3]{2}$
2. In each case, show that $u \in \mathbb{C}$ is algebraic over \mathbb{Q} and find the minimal polynomial.
 - (a) $u = \sqrt{2} + \sqrt{3}$
 - (b) $u = \sqrt[3]{2} + i$
 - (c) $u = \sqrt{1 + \sqrt{3}}$
 - (d) $u = \sqrt{1 + i}$
3. In each case, decide whether u is algebraic or transcendental over F and prove it.
 - (a) $u = \sqrt{\pi}$, $F = \mathbb{Q}(\pi)$
 - (b) $u = \sqrt{\pi}$, $F = \mathbb{Q}$
 - (c) $u = \pi^2$, $F = \mathbb{Q}$
 - (d) $u = 1 + \pi$, $F = \mathbb{Q}$
4. Show that u is algebraic over $F = \mathbb{Q}(v)$ and find the minimal polynomial if
 - (a) $u = 1 + i$, $v = \sqrt{2}$
 - (b) $u = \sqrt{2}$, $v = 1 + i$
5. If $u \in \mathbb{C}$, $u \notin \mathbb{R}$, show that $\mathbb{C} = \mathbb{R}(u)$.
6. If $E \supseteq F$ are fields, show that $F(u) = F(au)$ for all $u \in E$, $0 \neq a \in F$.

7. Find the minimal polynomial of $u = \sqrt{3} - i$: (a) over \mathbb{R} and (b) over \mathbb{Q} .
8. If $z = a + bi \in \mathbb{C}$, $a, b \in \mathbb{R}$, find the minimal polynomial of z over \mathbb{R} .
9. Show that $F(u, v) = F(u)$ if and only if $v = f(u)$ for some $f \in F[x]$.
10. If $u = \sqrt[3]{5}$, find the minimal polynomial of u over \mathbb{Q} and that of u^2 .
11. If $E \supseteq F$ are fields and $u \in E$, show that u is algebraic over F if and only if $[F(u) : F]$ is finite.
12. In each case, find a basis of E over \mathbb{Q} .
 - (a) $E = \mathbb{Q}(\sqrt[3]{2})$
 - (b) $E = \mathbb{Q}(1 - i)$
 - (c) $E = \mathbb{Q}(\sqrt{3}, \sqrt[3]{3})$
 - (d) $E = \mathbb{Q}(\sqrt{2}, \sqrt{3})$
 - (e) $E = \mathbb{Q}(\sqrt{3}, \sqrt{15})$
 - (f) $E = \mathbb{Q}(\sqrt{2}, \sqrt[3]{3})$
13. In each case, find $[E : F]$.
 - (a) $E = \mathbb{Q}(\sqrt{3} + \sqrt{5})$, $F = \mathbb{Q}(\sqrt{3})$
 - (b) $E = \mathbb{Q}(\sqrt{3}, \sqrt{15})$, $F = \mathbb{Q}(\sqrt{5})$
 - (c) $E = \mathbb{Q}(\sqrt{3} + i)$, $F = \mathbb{Q}(i)$
 - (d) $E = \mathbb{Q}(\sqrt[3]{3}, \sqrt{2})$, $F = \mathbb{Q}(\sqrt{2})$
14. Let $E \supseteq F$ be a finite extension and let $p \in F[x]$ be irreducible. If $p(u) = 0$ for some $u \in E$, show that $\deg p$ divides $[E : F]$.
15. If $E \supseteq F$ are fields and $[E : F]$ is prime, show that $E = F(u)$ for all $u \in E$, $u \notin F$.
16. If $E \supseteq F$ are fields and $u \in E$ has odd degree over F , show that $F(u) = F(u^2)$.
17. If $E \supseteq F$ are fields and $u \in E$ has prime degree over F , show that the only fields L such that $F(u) \supseteq L \supseteq F$ are $L = F$ and $L = F(u)$.
18. Let $E \supseteq L \supseteq F$ and $E \supseteq M \supseteq F$ be fields. If $[L : F]$ is prime, show that either $M \supseteq L$ or $M \cap L = F$.
19. Let $C \supseteq E \supseteq \mathbb{Q}$, where E is a field, and assume that $[E : \mathbb{Q}] = 2$. Show that $E = \mathbb{Q}(\sqrt{m})$, where m is a square-free integer.
20. Let $K \supseteq E \supseteq F$ be fields where $[E : F]$ is finite, and let $u \in K$ be algebraic over F .
 - (a) Show that $[E(u) : E] \leq [F(u) : F]$.
 - (b) Show that $[E(u) : F(u)] \leq [E : F]$.
21. Let $E \supseteq F$ be fields, and let $u, v \in E$ be algebraic over F of degrees m, n .
 - (a) Show that $[F(u, v) : F] \leq mn$.
 - (b) If m and n are relatively prime, show that $[F(u, v) : F] = mn$.
 - (c) Is the converse to (b) true? Support your answer.
22. If $E = F(u_1, \dots, u_n)$ and u_i is algebraic of degree m_i over F for each i , show that $[E : F] \leq m_1 m_2 \cdots m_n$.
23. Show that $\sqrt{2} \notin \mathbb{Q}(\pi)$. [Hint: Discussion following Lemma 1.]
24. Show that $[\mathbb{Q}(\pi) : \mathbb{Q}(\pi^3)]$ is finite and display a basis of $\mathbb{Q}(\pi)$ over $\mathbb{Q}(\pi^3)$.
25. Prove Lemma 1.
26. (a) If u^2 is algebraic over F , show that u is algebraic over F .
 - (b) If $f(u)$ is algebraic over F , $f \in F[x]$, $f \notin F$, show that u is algebraic over F .
27. Show that the intersection of any family of subfields of E is again a subfield of E .
28. Let $E \supseteq F$ be fields and let $u, v \in E$. If $u + v$ is algebraic over F , show that v is algebraic over $F(u)$. [Hint: Treat the case that u is transcendental separately.]
29. Is it possible that $u \notin F(v)$ is algebraic over $F(v)$ while v is transcendental over $F(u)$? Support your answer. [Hint: Exercise 23.]
30. Let $E \supseteq F$ be fields and let $u, v \in E$. If v is transcendental over F but algebraic over $F(u)$, show that u is algebraic over $F(v)$.
31. Let $E \supseteq F$ be fields and let $u \in E$ be transcendental over F .
 - (a) Show that $F(u) = \{f(u)g(u)^{-1} \mid f, g \in F[x]; g(x) \neq 0\}$.
 - (b) Show that $F(u) \cong F(x)$, the field of quotients of the integral domain $F[x]$.
 - (c) Show that every element $w \in F(u)$, $w \notin F$, is transcendental over F .

32. Let p and q in \mathbb{Q} satisfy $\sqrt{p} \notin \mathbb{Q}$ and $\sqrt{q} \notin \mathbb{Q}(\sqrt{p})$.
- Show that $\mathbb{Q}(\sqrt{p}, \sqrt{q}) = \mathbb{Q}(\sqrt{p} + \sqrt{q})$.
 - Use Theorem 5 to find a basis of $\mathbb{Q}(\sqrt{p}, \sqrt{q})$ over \mathbb{Q} .
 - Deduce that $x^4 - 2(p+q)x^2 + (p-q)^2$ is the minimal polynomial of $\sqrt{p} + \sqrt{q}$ over \mathbb{Q} .
33. Let $E \supseteq F$ be fields and let A be the algebraic closure of F in E . If $u \in E$, $u \notin A$, show that u is transcendental over A .
34. Let $E \supseteq F$ be fields and let $\{e_1, \dots, e_m\}$ be an F -basis of E . If $_E V$ is a vector space with basis $\{v_1, \dots, v_n\}$, show that $\dim(_F V) = mn$ and exhibit a basis of $_F V$.
35. (a) Let $E \supseteq R \supseteq F$, where $E \supseteq F$ is an algebraic extension of fields and R is a subring of E . Prove that R is a field.
- (b) Repeat (a) where R is an F -subspace of E , $\text{char } F \neq 2$, and $u \in R$ implies that $u^k \in R$ for all $k \geq 2$.
- (c) Show that (b) is false if $\text{char } F = 2$. [Hint: Let F be the quotient field of $\mathbb{Z}_2[x, y]$.]

6.3 SPLITTING FIELDS

So far our discussion of an algebraic element u over a field F has concerned a given extension field $E \supseteq F$ that contains u , and we have described the field $F(u)$ explicitly as a subfield of E . However, Theorem 4 §6.2 shows that

$$F(u) = \{a_0 + a_1 u + \cdots + a_{n-1} u^{n-1} \mid a_i \in F\} \cong \frac{F[x]}{\langle m \rangle},$$

where m is the (irreducible) minimal polynomial of u over F and $\deg m = n$. Hence, $F(u)$ does not depend on E , being completely determined by u and F .

So we change our perspective. Suppose that f is a nonconstant polynomial in $F[x]$ (possibly not irreducible), where F is a field. Is there a field $E \supseteq F$ in which f has a root? Leopold Kronecker solved this problem in the nineteenth century.

Theorem 1. Kronecker's Theorem. *If F is any field and f is a nonconstant polynomial in $F[x]$, there is an extension field of F in which f has a root.*

We proved this assertion earlier (Theorem 4 §4.3). The idea is simple. Because f has positive degree, it has a monic irreducible factor p . Hence, $E = F[x]/\langle p \rangle$ is a field (by Theorem 3 §4.3) that we explicitly describe as follows: We regard F as a subfield of E by identifying $a = a + \langle p \rangle$ for all $a \in F$. In addition, if we write $t = x + \langle p \rangle$, the elements of E take the form

$$E = \{a_0 + a_1 t + a_2 t^2 + \cdots + a_{n-1} t^{n-1} \mid a_i \in F\},$$

where $n = \deg p$. Moreover, $\{1, t, \dots, t^{n-1}\}$ is an F -basis of E , so $[E : F] = n$, where $n = \deg p$. Finally, $p(t) = 0$ (easily verified; see Lemma 4 §4.3), so $f(t) = 0$ because p is a factor of f . Thus, t is the desired root of f in E . Of course, $E = F(t)$ in the notation of Section 6.2, and p is the minimal polynomial of t over F . This not only proves Kronecker's theorem, but also provides a means of constructing field extensions of degree n over F by using monic irreducible polynomials of degree n .

Since every polynomial f is a product of irreducible factors, we can repeat this procedure and construct an extension of F over which f factors completely into linear factors. Here is an example where f is already irreducible.

Example 1. Find an extension $E \supseteq \mathbb{Z}_2$ in which $f = x^3 + x + 1$ factors completely into linear factors.

Solution. The polynomial f is irreducible over \mathbb{Z}_2 (it has no root in \mathbb{Z}_2) so

$$E = \{a_0 + a_1 t + a_2 t^2 \mid a_i \in \mathbb{Z}_2, f(t) = 0\}$$

is a field containing a root t of f by Theorem 2 §4.3. Hence, $x + t = x - t$ is a factor of f . The division algorithm gives $f = (x + t)g$, where $g = x^2 + tx + (1 + t^2)$, so it suffices to show that g also factors completely in E . Trial and error gives $g(t^2) = 0$, so $g = (x + t^2)(x + v)$ for some $v \in E$. Comparing coefficients of x gives $t = t^2 + v$, whence $v = t + t^2$. Thus, $g = (x + t^2)(x + t + t^2)$, and so $f = (x + t)(x + t^2)(x + t + t^2)$ factors completely in $E[x]$. \square

The following terminology is commonly used. Let f be a polynomial in $F[x]$ of degree $n \geq 1$, where F is a field. An extension field $E \supseteq F$ is called a **splitting field** of f over F if the following conditions are satisfied:

- (1) $f = a(x - u_1)(x - u_2) \cdots (x - u_n)$, $a \in F$, $u_i \in E$ for each i .
- (2) $E = F(u_1, u_2, \dots, u_n)$.

If (1) holds, we say that f splits over E or that f splits in $E[x]$. Hence, in Example 1, the field E is a splitting field of f over \mathbb{Z}_2 because $E = \mathbb{Z}_2(t)$.

If E is a splitting field of f over F , the only subfield of E (containing F) in which f splits is E itself. Indeed, u_1, \dots, u_n are the only roots of f in any subfield L of E containing F , so if f splits over L , then $L = E$ by (2).

Example 2. The field F is itself a splitting field of every linear polynomial in $F[x]$.

Example 3. The field $\mathbb{Q}(i)$ is a splitting field of $x^2 + 1$ over \mathbb{Q} because we have $x^2 + 1 = (x + i)(x - i)$ and $\mathbb{Q}(i, -i) = \mathbb{Q}(i)$.

For a nonconstant polynomial f in $\mathbb{Q}[x]$, the fundamental theorem of algebra asserts that f splits in $\mathbb{C}[x]$. If the roots are u_1, \dots, u_n , then $E = \mathbb{Q}(u_1, \dots, u_n)$ is a splitting field for f (which is contained in \mathbb{C}). The next theorem shows that splitting fields always exist (though they need not be subfields of \mathbb{C}).

Theorem 2. Let f be a polynomial of degree $n \geq 1$ over a field F . Then a splitting field $E \supseteq F$ of f over F exists and $[E : F] \leq n!$.

Proof. Use induction on $n \geq 1$. If $n = 1$, take $E = F$. If $n > 1$, let p be a monic irreducible factor of f , and by Kronecker's theorem, let $E \supseteq F$ be a field containing a root u_1 of p (and thus f). Put $E_1 = F(u_1)$ so that $u_1 \in E_1$ and $[E_1 : F] = \deg p \leq n$. Now $f = (x - u_1)g$ in $E_1[x]$, where $\deg g = n - 1$. Hence, by induction, let $E_2 \supseteq E_1$ be a splitting field for g with $[E_2 : E_1] \leq (n - 1)!$. Then $g = a(x - u_2) \cdots (x - u_n)$, where $a \in E_1$ and $u_i \in E_2$ for each i , so $E_2 = E_1(u_2, \dots, u_n) = F(u_1, u_2, \dots, u_n)$ is a splitting field for f over F . Finally,

$$[E_2 : F] = [E_2 : E_1][E_1 : F] \leq (n - 1)! \cdot n = n!$$

by the multiplication theorem, which completes the proof. \blacksquare

Example 4. Find a splitting field $E \supseteq \mathbb{Q}$ of $f = x^4 - 2x^2 - 3$, where $[E : \mathbb{Q}] = 4$.

Solution. Because $f = (x^2 - 3)(x^2 + 1)$, the roots of f in \mathbb{C} are $\pm\sqrt{3}$ and $\pm i$. Hence, $E = \mathbb{Q}(\sqrt{3}, i)$ is the required splitting field. We have $E = L(i)$, where $L = \mathbb{Q}(\sqrt{3})$ and $[E : L] = 2 = [L : \mathbb{Q}]$, as the reader can verify. Hence, $[E : \mathbb{Q}] = 2 \cdot 2 = 4$ by the multiplication theorem. \square

Example 4 shows that if a polynomial f has degree n over F , the degree over F of its splitting field can be much smaller than the bound of $n!$ in Theorem 2. Nonetheless, this bound is the best possible, as Example 5 shows.

Example 5. Find a splitting field E of $f = x^3 - 5$ over \mathbb{Q} such that $[E : \mathbb{Q}] = 6$.

Solution. The roots of f in \mathbb{C} are u, uw , and uw^2 , where $u = \sqrt[3]{5}$, and $w = e^{2\pi i/3}$ is a cube root of unity. Clearly, $E = \mathbb{Q}(u, uw, uw^2) = \mathbb{Q}(u, w)$ is a splitting field. Write $L = \mathbb{Q}(u)$. As f is irreducible over \mathbb{Q} , it is the minimal polynomial of u , so $[L : \mathbb{Q}] = 3$. Since $[E : \mathbb{Q}] = [E : L][L : \mathbb{Q}]$, it remains to show that $[E : L] = 2$. As $E = L(w)$, this follows if we can show that the minimal polynomial of w over L has degree 2. Note that

$$f = x^3 - u^3 = (x - u)(x^2 + ux + u^2)$$

in $L[x]$. If we write $p = x^2 + ux + u^2$, then $p(w) = 0$ because $f(w) = 0$ and $w \neq u$. Similarly, $p(w^2) = 0$, so p has no root in L ($L \subseteq \mathbb{R}$ but $w \notin \mathbb{R}$ and $w^2 \notin \mathbb{R}$). Thus, p is irreducible over L , and so it is the minimal polynomial of w . Hence, $[E : L] = \deg p = 2$, which completes the argument. \square

Example 6. If F is any field, show that any splitting field of a quadratic f in $F[x]$ is a simple extension $F(u)$ of F .

Solution. Let $f = ax^2 + bx + c$, $a \neq 0$, and let $E \supseteq F$ be a splitting field of f . If $u, v \in E$ are the roots of f , then $f = a(x - u)(x - v)$, so comparing coefficients of x gives $b = -a(u + v)$. Thus $v = -u - a^{-1}b \in F(u)$, so $E = F(u, v) = F(u)$. \square

Example 7. Two different irreducible polynomials can have the same splitting field. For example, both $x^2 - 2$ and $x^2 - 2x - 1$ have splitting field $\mathbb{Q}(\sqrt{2})$ because the roots are $\pm\sqrt{2}$ and $1 \pm \sqrt{2}$, respectively.

Example 7 notwithstanding, the splitting field of a polynomial f in $F[x]$ is uniquely determined by f up to isomorphism. In fact, we prove a slightly stronger result that utilizes a commonly occurring concept in field theory.

Let $R \supseteq F$ and $\bar{R} \supseteq \bar{F}$, where F and \bar{F} are fields that are subrings of the rings R and \bar{R} , respectively. Given a ring isomorphism $\sigma : F \rightarrow \bar{F}$, a ring isomorphism $\hat{\sigma} : R \rightarrow \bar{R}$ is said to extend σ if $\hat{\sigma}(a) = \sigma(a)$ holds for every $a \in F$ (see the diagram).

$$\begin{array}{ccc} R & \xrightarrow{\hat{\sigma}} & \bar{R} \\ | & & | \\ F & \xrightarrow{\sigma} & \bar{F} \end{array}$$

An important instance occurs in the following context. Let $\sigma : F \rightarrow \bar{F}$ be an isomorphism of fields. Given $f \in F[x]$, define a new polynomial $f^\sigma \in \bar{F}[x]$ as follows: If $f = a_0 + a_1x + \cdots + a_nx^n$, $a_i \in F$, let

$$f^\sigma = \sigma(a_0) + \sigma(a_1)x + \cdots + \sigma(a_n)x^n. \quad (*)$$

Then the mapping $F[x] \rightarrow \bar{F}[x]$ given by $f \mapsto f^\sigma$ is a ring isomorphism that extends σ (Exercise 16). This mapping is very useful; our present interest in it is as follows.

Suppose p is a monic irreducible polynomial in $F[x]$. Then p^σ is monic in $\bar{F}[x]$ (as $\sigma(1) = 1$) and irreducible (because $\deg f^\sigma = \deg f$ for all $f \in F[x]$). Now define

$$\varphi : \frac{F[x]}{\langle p \rangle} \rightarrow \frac{\bar{F}[x]}{\langle p^\sigma \rangle} \quad \text{by} \quad \varphi(f + \langle p \rangle) = f^\sigma + \langle p^\sigma \rangle \quad (**)$$

for all $f \in F[x]$. Then φ is well defined (and one-to-one) because

$$\begin{aligned} f + \langle p \rangle = g + \langle p \rangle &\Leftrightarrow p|(f - g) \text{ in } F[x] \\ &\Leftrightarrow p^\sigma|(f^\sigma - g^\sigma) \text{ in } \bar{F}[x] \\ &\Leftrightarrow f^\sigma + \langle p^\sigma \rangle = g^\sigma + \langle p^\sigma \rangle. \end{aligned}$$

Now the fact that $f \mapsto f^\sigma$ is a ring isomorphism shows that φ is also a ring isomorphism. We need this result in the proof of Theorem 3.

Theorem 3. *Let $\sigma : F \rightarrow \bar{F}$ be an isomorphism of fields. Given a monic irreducible polynomial p in $F[x]$, let u be a root of p in an extension field $E \supseteq F$ and let v be a root of p^σ in an extension field $\bar{E} \supseteq \bar{F}$. Then there is a unique isomorphism*

$$F(u) \rightarrow \bar{F}(v) \quad \text{given by} \quad f(u) \mapsto f^\sigma(v), \quad f \in F[x]$$

that extends σ and carries u to v .

Proof. The polynomial p is the minimal polynomial of u over F . Hence, as in the proof of Theorem 4 §6.2, the mapping $\theta : F[x] \rightarrow F(u)$ given by $\theta(f) = f(u)$ is an onto ring homomorphism with $\ker \theta = \langle p \rangle$. This mapping induces an isomorphism $F(u) \cong F[x]/\langle p \rangle$ given by $f(u) \leftrightarrow f + \langle p \rangle$. Similarly, $\bar{F}(v) \cong \bar{F}[x]/\langle p^\sigma \rangle$. Now compose these mappings with the isomorphism φ in $(**)$ to get

$$\begin{aligned} F(u) &\rightarrow \frac{F[x]}{\langle p \rangle} \xrightarrow{\varphi} \frac{\bar{F}[x]}{\langle p^\sigma \rangle} \rightarrow \bar{F}(v), \\ f(u) &\mapsto f + \langle p \rangle \mapsto f^\sigma + \langle p^\sigma \rangle \mapsto f^\sigma(v). \end{aligned}$$

Hence, the composite map $f(u) \mapsto f^\sigma(v)$ is an isomorphism $F(u) \rightarrow \bar{F}(v)$. This map carries u to v (take $f = x$), so it remains to verify that it extends σ . If $a \in F$, let $g = a$ be the corresponding constant polynomial. Then $g^\sigma(x) = \sigma(a)$, so $a = g(u) \mapsto g^\sigma(v) = \sigma(a)$, as required. The uniqueness is Exercise 16. ■

The special case of Theorem 3 where $\bar{F} = F$ is worth mentioning. Let p be a monic irreducible polynomial and let u and v be two roots of p in suitable extension fields $E \supseteq F$ and $\bar{E} \supseteq F$. Then the fields $F(u)$ and $F(v)$ are isomorphic. In fact, if $\deg p = n$, the map $\hat{\sigma} : F(u) \rightarrow F(v)$ given by

$$\hat{\sigma}(a_0 + a_1 u + \cdots + a_{n-1} u^{n-1}) = a_0 + a_1 v + \cdots + a_{n-1} v^{n-1}$$

is an isomorphism that carries u to v and fixes F in the sense that $\hat{\sigma}(a) = a$ for all $a \in F$ (that is, $\hat{\sigma}$ extends the identity map $1_F : F \rightarrow F$).

Theorem 3 will be used repeatedly in Chapter 10. For now, our chief interest in the result is that it enables us to prove that splitting fields are unique.

Theorem 4. *Let $\sigma : F \rightarrow \bar{F}$ be an isomorphism of fields, let $f \in F[x]$ be a nonconstant polynomial, and let $f^\sigma \in \bar{F}[x]$ be as given in $(*)$. If $E \supseteq F$ is a splitting field for f and $\bar{E} \supseteq \bar{F}$ is a splitting field for f^σ , there is an isomorphism $E \rightarrow \bar{E}$ that extends σ .*

Proof. Use induction on $n = \deg f = \deg f^\sigma$ (see the diagram). If $n = 1$, then $E = F$ and $\bar{E} = \bar{F}$, so σ itself is the required map. If $n > 1$, let $u \in E$ be a root of a monic irreducible divisor p of f and let $v \in \bar{E}$ be a root of p^σ . Then σ extends (by Theorem 3) to an isomorphism $\tau : F(u) \rightarrow \bar{F}(v)$ such that $\tau(u) = v$. Now write $f = a(x - u)(x - u_1) \dots (x - u_n)$ in $E[u]$, $a \in F$, $u_i \in E$, $v_i \in \bar{E}$, and define $g = a(x - u_1) \dots (x - u_n)$. Then $E = F(u)(u_2 \dots u_n)$ is a splitting field of g over $F(u)$. However, $f^\sigma = f^\tau = [x - \tau(u)]g^\tau = (x - v)g^\tau$, and hence \bar{E} is a splitting field of g^τ over $\bar{F}(v)$. Then, by induction, there is an isomorphism $E \rightarrow \bar{E}$ that extends τ and so extends σ . This completes the proof. ■

$$\begin{array}{ccc} E & \xrightarrow{\quad} & \bar{E} \\ | & & | \\ F(u) & \xrightarrow{\quad \tau \quad} & \bar{F}(v) \\ | & & | \\ F & \xrightarrow{\quad \sigma \quad} & \bar{F} \end{array}$$

Example 8. Find a splitting field for $f = x^4 + x^3 + x^2 + 1$ in $\mathbb{Z}_2[x]$, and factor f completely in $\mathbb{Z}_2[x]$.

Solution. We have $f(1) = 0$ because $\text{char } \mathbb{Z}_2 = 2$, so $f = (x + 1)(x^3 + x + 1)$ by the division algorithm. Now $g = x^3 + x + 1$ is irreducible over \mathbb{Z}_2 (it has no root in \mathbb{Z}_2) so, as in Example 7 §4.3, we obtain a field

$$E = \{a_0 + a_1t + a_2t^2 \mid a_i \in \mathbb{Z}_2 \text{ and } t^3 = 1 + t\}.$$

Then t is a root of g , so $g = (x + t)(x^2 + tx + (1 + t^2))$, again by the division algorithm. Now it remains to determine if $h = x^2 + tx + (1 + t^2)$ splits in $E[x]$. If we set $h(a + bt + ct^2) = 0$, where $a, b, c \in \mathbb{Z}_2$, then comparing coefficients of x (and using the fact that $d^2 = d$ for any $d \in \mathbb{Z}_2$), we obtain the roots t^2 and $t + t^2$. Hence, $h = (x + t^2)(x + t + t^2)$ and we obtain $f = (x + 1)(x + t)(x + t^2)(x + t + t^2)$, a complete factorization in $E[x]$. In particular, E is a splitting field for f . □

Algebraic Closures

In view of all our efforts to find splitting fields for polynomials, the fact that every nonconstant polynomial in $C[x]$ splits is remarkable, to say the least. The next theorem characterizes when this happens.

Theorem 5. *The following conditions on a field C are equivalent:*

- (1) *Every nonconstant polynomial in $C[x]$ has a root in C .*
- (2) *Every irreducible polynomial in $C[x]$ has degree 1.*
- (3) *Every nonconstant polynomial in $C[x]$ splits in $C[x]$.*
- (4) *If $E \supseteq C$ is an algebraic extension, then $E = C$.*

Proof. (1) \Rightarrow (2) \Rightarrow (3). These are left to the reader.

(3) \Rightarrow (4). If $u \in E$, let $f(u) = 0$, where f is a nonzero polynomial in $C[x]$. Then f is not constant, so by (3), $f = a(x - b_1)(x - b_2) \dots (x - b_n)$, where $a, b_i \in C$. Thus, $f(u) = 0$ means that $u = b_i$ for some i , so $u \in C$ proving (4).

(4) \Rightarrow (1). If f is a nonconstant polynomial in $C[x]$, let u be a root of f in some extension field E by Theorem 1. Thus, $C(u) \supseteq C$ is an algebraic extension (it is finite by Theorem 4 §6.2), so $C(u) = C$ by (4). Hence, $u \in C$ and (1) follows. ■

We can express condition (4) in Theorem 5 by saying that the field C has no proper algebraic extension. With this in mind, we say that a field C is **algebraically closed** if it satisfies the conditions in Theorem 5. The fundamental theorem of algebra (Section 6.6) asserts that each polynomial in $\mathbb{C}[x]$ has a root in \mathbb{C} . Thus:

Example 9. \mathbb{C} is algebraically closed.

If $E \supseteq F$ is a field extension, Corollary 2 of Theorem 6 §6.2 shows that

$$A = \{u \in E \mid u \text{ is algebraic over } F\}$$

is a subfield of E containing F , called the algebraic closure of F in E . The field A is clearly the largest algebraic extension of F contained in E . If E is algebraically closed, we have Theorem 6.

Theorem 6. *If $C \supseteq F$ are fields and C is algebraically closed, then the algebraic closure*

$$A = \{u \in C \mid u \text{ is algebraic over } F\}$$

of F in C is itself algebraically closed.

Proof. If f is a nonconstant polynomial in $A[x]$, then $f \in C[x]$, so f has a root u in C by hypothesis. By Theorem 5, we must show that $u \in A$. If $f = a_0 + a_1x + \dots + a_nx^n$, $a_i \in A$, write $E = F[a_0, a_1, \dots, a_n]$. Then $[E : F]$ is a finite extension (by Theorem 6 §6.2 because each a_i is algebraic over F), and $[E(u) : E]$ is finite because u is algebraic over E . Hence, $E(u) \supseteq F$ is finite and so algebraic. Since $u \in E(u)$, it follows that u is algebraic over F as required. ■

If F is a field, a field extension $A \supseteq F$ is called an **algebraic closure** of F if A is an algebraic extension of F that is algebraically closed. Thus, an algebraic closure of a field F is a maximal algebraic extension of F in the sense that it is an algebraic extension with no proper algebraic extension (by Theorem 5). Theorem 6 shows that any subfield F of an algebraically closed field C has an algebraic closure (its algebraic closure in C). Recall that the field \mathbb{A} of algebraic numbers is defined by

$$\mathbb{A} = \{u \in \mathbb{C} \mid u \text{ is algebraic over } \mathbb{Q}\}.$$

Then specializing Theorem 6 to the case $F = \mathbb{Q}$, $C = \mathbb{C}$, gives

Corollary. *The field \mathbb{A} of algebraic numbers is an algebraic closure of \mathbb{Q} .*

In fact, we have Theorem 7.

Theorem 7. *Every field F has an algebraic closure $A \supseteq F$. Moreover, if $A' \supseteq F$ is another algebraic closure, there is an isomorphism $\sigma: A \rightarrow A'$ that fixes F .*

The proof requires Zorn's Lemma (see Appendix C) and is omitted.⁷²

Leopold Kronecker (1823–1891) Kronecker was born into a prosperous family and in addition to his mathematical studies, he actively pursued business interests in his early years. He was so successful that by the time he was 30, he could afford to devote himself entirely to mathematics. He eventually succeeded his teacher Kummer as professor at the University of Berlin.

⁷²See McCarthy, P.J., *Algebraic Extensions of Fields*, Waltham, MA: Blaisdell, 1966, p. 22.

Kronecker worked primarily in algebraic number theory, and he is said to be one of the inventors of the theory (with Kummer and Dedekind). He produced mathematics of first quality and was one of the first algebraists to understand thoroughly the work of Galois. However, he insisted on dealing only with numbers such as $\sqrt{2}$, which he could construct from the rational numbers by a finite process. He categorically rejected the real-number constructions of his day that, using infinite limiting processes, gave meaning to transcendental numbers such as π . He used to say, "God made the integers, and all the rest is the work of man."

This point of view brought him into conflict with Karl Weierstrass and Georg Cantor who were creating modern analysis and set theory. While they remained friends, Kronecker and Weierstrass debated this issue all their lives. However, Kronecker's attack deeply affected the hypersensitive Cantor and was likely a factor in the breakdowns he suffered in his later years. Cantor subsequently was awarded the recognition he deserved, whereas Kronecker's point of view found little support among mathematicians of the day.

Exercises 6.3

Throughout these exercises, F denotes a field.

1. In each case, find the splitting field E of f over \mathbb{Q} and find $[E : \mathbb{Q}]$.
 - (a) $f = x^3 + 1$
 - (b) $f = x^4 + 1$
 - (c) $f = x^4 - 6x^2 - 7$
 - (d) $f = x^6 + 2x^3 - 3$
2. (a) Find the splitting field of $f = x^4 - 2x^3 - 7x^2 + 10x + 10$ over \mathbb{Q} .
 (b) Find the splitting field of $f = x^4 + x^3 + 2x^2 + x + 1$ over \mathbb{Q} .
3. If $2 \neq 0$ in the field F , show that the splitting field E of $x^4 + 1$ over F is a simple extension of F and factors $x^4 + 1$ completely in $E[x]$. What happens if $2 = 0$ in F ?
4. In each case, find the splitting field E of f over F and factor f completely in E .
 - (a) $f = x^3 + 1, F = \mathbb{Z}_2$
 - (b) $f = x^3 + 1, F = \mathbb{Z}_3$
 - (c) $f = x^3 + x^2 + 1, F = \mathbb{Z}_2$
 - (d) $f = x^3 - x + 1, F = \mathbb{Z}_3$
 - (e) $f = x^4 - x^2 - 2, F = \mathbb{Z}_3$
 - (f) $f = x^4 + x^3 + x + 1, F = \mathbb{Z}_2$
5. Show that $x^2 - 3$ and $x^2 - 2x - 2$ have the same splitting field.
6. (a) Is \mathbb{C} the splitting field of some polynomial over \mathbb{Q} ? Support your answer.
 (b) If $f \in \mathbb{R}[x]$ is nonconstant, show that \mathbb{R} or \mathbb{C} is a splitting field of f over \mathbb{R} .
7. Let $f = gh$ in $F[x]$ where g and h are nonconstant. If E is a splitting field of f over F , show that g splits in $E[x]$.
8. Let $E \supseteq F$ be a splitting field of f over F . If $[E : F]$ is prime, show that $E = F(u)$ for some u in E (that is, E is a simple extension of F).
9. Let f and g be polynomials in $F[x]$. Show that f and g are relatively prime in $F[x]$ if and only if they have no common root in any extension $E \supseteq F$.
10. If $f \in F[x]$, show that $E \supseteq F$ is a splitting field for f if and only if f splits over E and not over any proper subfield of E containing F .
11. Let $E \supseteq L \supseteq F$ be fields and let $f \in F[x]$. If E is a splitting field for f over F , show that E is also a splitting field of f over L .
12. If $f, g \in F[x]$, show that any splitting field of fg contains splitting fields of f and g .
13. Let $w = e^{2\pi i/p}$ be a p th root of unity, where p a prime. Show that $\mathbb{Q}(w)$ is the splitting field of $x^p - 1$ over \mathbb{Q} and that $[\mathbb{Q}(w) : \mathbb{Q}] = p - 1$. [Hint: Example 13 §4.2.]
14. Let $f \in F[x]$ and let $g = f(ax + b)$, $a \neq 0$, b in F . From Exercise 12, assume that $K \supseteq F$ is a field containing splitting fields E and L of f and g , respectively. Show that $E = L$.

15. Let f and g be monic and irreducible in $F[x]$ with relatively prime degrees. If u is a root of g in some extension field $E \supseteq F$, show that f is irreducible over $F(u)$. [Hint: Use Theorem 1 to find a field $K \supseteq E$ in which f has a root v . Apply Exercise 21 §6.2 to show that f is the minimal polynomial of v over $F(u)$.]
16. Show that the isomorphism $F(u) \rightarrow \bar{F}(v)$ in Theorem 3 is uniquely determined by the condition that it extends σ and carries $u \mapsto v$.
17. If $\sigma : F \rightarrow \bar{F}$ is an isomorphism of fields, prove that $f \mapsto f^\sigma$ is a ring isomorphism $F[x] \rightarrow \bar{F}[x]$ that extends σ (see the discussion preceding Theorem 3).
18. If $E \supseteq F$ is an algebraic extension of fields and every polynomial in $F[x]$ splits over E , show that E is algebraically closed. [Hint: Corollary 1 of Theorem 6 §6.2.]
19. Show that π is not algebraic over the field A of algebraic numbers.
20. (a) Find the algebraic closure A of \mathbb{Q} in $E = \mathbb{Q}(i, \pi)$. [Hint: Exercises 19 and 31 §6.2.]
(b) Is A algebraically closed? Support your answer.
21. Show that the following conditions are equivalent for fields $E \supseteq F$:
- (1) E is the splitting field of a polynomial in $F[x]$.
 - (2) $[E : F]$ is finite and every irreducible polynomial in $F[x]$ with a root in E splits completely in $E[x]$.
- Algebraic extensions with the second property in (2) are called **normal extensions**.
[Hint: For (1) \Rightarrow (2), let p in $F[x]$ be irreducible with a root u in E and let v be a root of p in a field $K \supseteq E$. Find an isomorphism $\sigma : F(u) \rightarrow F(v)$ and apply Theorem 3 to conclude that $E \cong E(v)$. Then argue that $E = E(v)$, so $v \in E$. For (2) \Rightarrow (1), use Theorem 6 §6.2.]

$$\begin{array}{ccc} & E & \\ & | & \\ F(u) & \xrightarrow{\sigma} & F(v) \\ & \searrow & \swarrow \\ & F & \end{array}$$

6.4 FINITE FIELDS

The theory of finite fields is satisfying because they can be completely classified. Galois introduced this subject in his investigation of the insolvability of polynomial equations. Apart from its intrinsic interest, the subject has applications in group theory, combinatorics, and coding theory, among other areas. Of course, when we speak of a finite field F , we mean that its **order** $|F|$ is finite.

If F is a finite field, our first observation is that F has characteristic p for some prime p . Therefore, F contains a copy of the field \mathbb{Z}_p of integers modulo p . It is customary (and we shall do so) to identify \mathbb{Z}_p with the prime subfield of F , that is, $\mathbb{Z}_p \subseteq F$. In particular, this means that F is a vector space over \mathbb{Z}_p and so has a basis $\{u_1, \dots, u_n\}$. Thus, the elements of F are uniquely represented in the form $a_1u_1 + \dots + a_nu_n$, $a_i \in \mathbb{Z}_p$, by Theorem 3 §6.1. There are p independent choices for each coefficient a_i , so we have

Theorem 1. *If F is a finite field, then $|F| = p^n$ for some $n \geq 1$, where $p = \text{char } F$.*

Theorem 1 leads inevitably to two questions: Is there a field of order p^n for each prime p and integer $n \geq 1$? If so, is it unique? The answer to both questions is yes (Theorem 4). One method of constructing a field of p^n elements is already available: If f is an irreducible polynomial of degree n in $\mathbb{Z}_p[x]$, then the factor ring $\mathbb{Z}_p[x]/\langle f \rangle$ is a field of order p^n by Theorem 3 §4.3. The problem is that we have no guarantee that such a polynomial f exists.

To motivate the procedure we use, suppose for the moment that a field F exists with $|F| = p^n$ elements. Then F^* is a group with $p^n - 1$ elements so, by Lagrange's theorem, $a^{p^n-1} = 1$ for all $a \neq 0$ in F . Hence, $a^{p^n} = a$, so as this also holds if $a = 0$, every element of F is a root of the polynomial $x^{p^n} - x$. Hence, the approach we take is to show that the splitting field of $x^{p^n} - x$ over \mathbb{Z}_p is a field of order p^n . This method has the added virtue that the uniqueness of the field then comes from the uniqueness theorem for splitting fields (Theorem 4 §6.3). Moreover, we can then prove the existence of an irreducible polynomial of each degree over \mathbb{Z}_p .

The construction of the splitting field of $x^{p^n} - x$ requires two preliminary observations. The first is related to the binomial theorem: If F is any field of characteristic p , and if a and b are elements of F , then

$$(a + b)^p = a^p + b^p \quad \text{for all } a, b \in F$$

by Theorem 2 §3.4.⁷³ Thus, the mapping $\sigma : F \rightarrow F$ given by $\sigma(a) = a^p$ is a ring homomorphism. It is one-to-one because F is a field, and so is onto because F is finite. Thus, σ is an automorphism, called the **Frobenius automorphism** of F .

The second result we need to compute the splitting field of $x^{p^n} - x$ is a condition guaranteeing that a polynomial in $F[x]$ has distinct roots in any splitting field. This requires a purely algebraic version of the derivative of a polynomial.

Let $f = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ be a polynomial in $F[x]$. The **derivative** of f is the polynomial f' in $F[x]$ defined by

$$f' = a_1 + 2a_2x + \cdots + na_nx^{n-1}.$$

In particular, if $f = ax^k$, where $k \geq 0$, then $f' = kax^{k-1}$. This relation holds even if $k = 0$, because the derivative of a constant polynomial is 0. Note that this definition of the derivative does not involve limits as in calculus. Nonetheless, the usual rules of differentiation hold.

Theorem 2. Let f and g be polynomials in $F[x]$, where F is a field.

- (1) $(af)' = af'$ for all $a \in F$.
- (2) $(f + g)' = f' + g'$.
- (3) $(fg)' = f'g + fg'$.
- (4) $\{f[g(x)]\}' = f'[g(x)]g'(x)$.

Proof. (1) and (2) follow immediately from the definition. To prove (3), write $f = a_0 + a_1x + a_2x^2 + \cdots$. If y is an indeterminate over $F[x]$, compute

$$\begin{aligned} f(x) - f(y) &= a_1(x - y) + a_2(x^2 - y^2) + a_3(x^3 - y^3) + \cdots \\ &= (x - y)[a_1 + a_2(x + y) + a_3(x^2 + xy + y^2) + \cdots] \\ &= (x - y)f_0(x, y), \end{aligned}$$

where $f_0(x, y)$ is a polynomial uniquely determined by f . Our interest in this is the observation that $f_0(x, x) = f'$. Compute in the field of quotients of $F[x, y]$:

$$\begin{aligned} p_0(x, y) &= \frac{p(x) - p(y)}{x - y} = f \left[\frac{g(x) - g(y)}{x - y} \right] + \left[\frac{f(x) - f(y)}{x - y} \right] g \\ &= f g_0(x, y) + f_0(x, y) g. \end{aligned}$$

Now (3) follows by taking $y = x$, and a similar argument gives (4) (Exercise 16). ■

⁷³This formula is sometimes called the “freshman’s dream.”

Thus, we can compute derivatives of polynomials over any field just as we do over \mathbb{R} in calculus.

If $f \in F[x]$, F a field, an element a of F is called a **repeated root** of f if $f = (x - a)^2 g$ for some $g \in F[x]$. Here is a simple test for the existence of repeated roots.

Theorem 3. Let f be a polynomial in $F[x]$, F a field, and let $a \in F$. Then:

- (1) $(x - a)^2$ divides f if and only if $(x - a)$ divides both f and f' .
- (2) If $f(a) = 0$ then $(x - a)^2$ divides f if and only if $f'(a) = 0$.

Proof. If $f = (x - a)^2 g$, then $f' = (x - a)[(x - a)g' + 2g]$ by Theorem 2. Conversely, if $f = (x - a)h$, then $f' = (x - a)h' + h$. Thus, $(x - a)$ divides h because $x - a$ divides f' , so $(x - a)^2$ divides f . This proves (1), and (2) is now clear. ■

We can now prove the main theorem of this section.

Theorem 4. Let p be prime, let $n \geq 1$ be an integer, and write $f = x^{p^n} - x$.

- (1) Any field F with $|F| = p^n$ is a splitting field of f over \mathbb{Z}_p .
- (2) Every splitting field of f over \mathbb{Z}_p has order p^n .

Hence, a field of order p^n exists and is unique up to an isomorphism fixing \mathbb{Z}_p .

Proof. (1) Assume that $|F| = p^n$. We observed above that every element of F is a root of f . As $\deg f = p^n$, f can have at most p^n roots in any field. Thus, the fact that $|F| = p^n$ implies that f factors into linear polynomials in $F[x]$. Hence, F is a splitting field for f .

(2) Let K be a splitting field of f over \mathbb{Z}_p and let $K_0 = \{a \in K \mid f(a) = 0\}$ denote the set of roots of f in K . We have $f' = -1 \neq 0$, so f has distinct roots in K by Theorem 3. Because f splits in K and $\deg f = p^n$, this implies that $|K_0| = p^n$. Hence, it suffices to show that K_0 is a subfield of K (then $K_0 = K$ because K is generated by the roots of f). To this end, let $\sigma : K \rightarrow K$ be the Frobenius automorphism given by $\sigma(a) = a^p$. Then $\sigma^2(a) = \sigma(a^p) = \sigma(a)^p = a^{p^2}$ and an easy induction gives $\sigma^n(a) = a^{p^n}$. This means that $K_0 = \{a \in K \mid \sigma^n(a) = a\}$. Because σ^n is an automorphism of K , it follows that K_0 is a subfield of K , as required.

Finally, the existence of a field of order p^n follows from (2) and Theorem 2 §6.3. The uniqueness is by Theorem 4 §6.3. ■

If p is a prime and $n \geq 1$ is an integer, the unique field with p^n elements is called the **Galois field** of order p^n and is denoted $GF(p^n)$.

Example 1. $GF(p) = \mathbb{Z}_p$ for each prime p .

We have already constructed the Galois fields $GF(4)$ and $GF(8)$ (Example 7 §4.3 and Example 1 §6.3), by using the fact that $x^2 + x + 1$ and $x^3 + x + 1$ are irreducible over \mathbb{Z}_2 . The polynomial $x^2 + 1$ is irreducible over \mathbb{Z}_p for any prime p congruent to 3 modulo 4 (Example 6 §4.3), which yields $GF(p^2)$ in this case. However, finding an irreducible polynomial of degree p^n over \mathbb{Z}_p is not easy (although we will show in Corollary 2, Theorem 7, that one must exist).

Example 2. Show that $x^4 + x + 1$ is irreducible over \mathbb{Z}_2 and so construct $GF(16)$.

Solution. It suffices to show that $f = x^4 + x + 1$ is irreducible; then Theorem 3 §4.3 gives $GF(16) = \{a + bt + ct^2 + dt^3 \mid a, b, c, d \in \mathbb{Z}_2; t^4 = t + 1\}$. Suppose that f is not

irreducible. Because f has no root in \mathbb{Z}_2 , it must factor as $f = pq$, where p and q are quadratics in $\mathbb{Z}_2[x]$. But then $p = q = x^2 + x + 1$ (the other quadratics are x^2 , $x^2 + 1$, and $x^2 + x$, and all have a root in \mathbb{Z}_2). Hence, $f = pq = (x^2 + x + 1)^2 = x^4 + x^2 + 1$, a contradiction. \square

If G is a cyclic group of order n , G has a subgroup of order m if and only if $m|n$ and, in this case, there is exactly one subgroup of order m (Theorem 9 §2.4). However, the problem of describing all subgroups of an arbitrary finite group G is very difficult (although much can be said if G is abelian). In the case of finite fields, however, we can describe the subfields of $GF(p^n)$ explicitly.

Theorem 5. Let p be prime and let $n \geq 1$ be an integer.

- (1) If K is a subfield of $GF(p^n)$, then $K \cong GF(p^m)$ for some m with $m|n$.
- (2) If $m|n$, there is exactly one subfield of $GF(p^n)$ of order p^m , and it consists of the roots of $x^{p^m} - x$ in $GF(p^n)$.

Proof. (1) Write $E = GF(p^n)$. Given a subfield $K \subseteq E$, then $\text{char } K = p$, so $\mathbb{Z}_p \subseteq K$. Also, $K \cong GF(p^m)$ for some $m \leq n$ by Theorem 1. In fact, $m|n$ by the multiplication theorem: $n = [E : \mathbb{Z}_p] = [E : K][K : \mathbb{Z}_p] = [E : K]m$.

(2) Observe that $x^{ab} - 1 = (x^a - 1)(x^{ab-a} + x^{ab-2a} + \dots + x^a + 1)$. Consequently, if $n = mk$, we have $p^n - 1 = (p^m - 1)q$ for some $q \in \mathbb{Z}$. Hence,

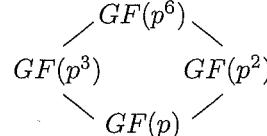
$$x^{p^n} - x = x(x^{p^n-1} - 1) = x(x^{p^m-1} - 1)g = (x^{p^m} - x)g,$$

where $g \in F[x]$. Hence, $(x^{p^m} - x)$ splits in $E[x]$ since $x^{p^n} - x$ does. If we define $E_0 = \{u \in E \mid u \text{ is a root of } x^{p^m} - x\}$, then $|E_0| = p^n$ because the roots of $(x^{p^m} - x)$ are distinct. Moreover, E_0 is a field as in the proof of Theorem 4.

Now let $K \subseteq E$ be any subfield with $|K| = p^m$. Then $K \subseteq E_0$ by Theorem 4. But then $K = E_0$ because $|K| = |E_0|$, proving (2). \blacksquare

Example 3. Draw the lattice diagram of the subfields of $GF(p^6)$.

Solution. By Theorem 5, the subfields are $GF(p) = \mathbb{Z}_p$, $GF(p^2)$, $GF(p^3)$, and $GF(p^6)$. The lattice diagram is shown at the right. \square



If $f \in \mathbb{Z}[x]$ is monic, the modular irreducibility test (Theorem 7 §4.2) asserts that if f is irreducible in $\mathbb{Z}_p[x]$ for some prime p , then f is irreducible in $\mathbb{Q}[x]$; that is, it has no proper factorization in $\mathbb{Z}[x]$. But the converse is false by Theorem 5:

Example 4. Show that $x^4 + 1$ is irreducible in $\mathbb{Q}[x]$, but is reducible in $\mathbb{Z}_p[x]$ for every prime p .

Solution. By Theorem 9 §4.1, $x^4 + 1$ has no root in \mathbb{Q} . If $f = (x^2 + ax + b)(x^2 + cx + d)$ in $\mathbb{Z}[x]$, then $a + c = 0$, $b + ac + d = 0$ and $bd = 1$, so $a^2 = \pm 2$, a contradiction. So $x^4 + 1$ is irreducible in $\mathbb{Q}[x]$.

Write $E = GF(p^2)$ for convenience, and regard $\mathbb{Z}_p \subseteq E$. Suppose that $x^4 + 1$ is irreducible over \mathbb{Z}_p . Then p is odd (otherwise $x^4 + 1 = (x^2 + 1)^2$, so $8|(p^2 - 1)$). Hence $x^8 - 1$ divides $x^{p^2-1} - 1$ in $\mathbb{Z}[x]$ so, since $x^8 - 1 = (x^4 + 1)(x^4 - 1)$, we have $x^{p^2-1} - 1 = (x^4 + 1)q$ for some $q \in \mathbb{Z}[x]$. But then $(x^4 + 1)q$ has $p^2 - 1$ distinct roots in E by Theorem 5, and it follows that $x^4 + 1$ has a root $u \in E \setminus \mathbb{Z}_p$. Hence $[\mathbb{Z}_p(u) : \mathbb{Z}_p] \leq 4 < p^2$ so $E \supset \mathbb{Z}_p(u)$. Thus $E \supset \mathbb{Z}_p(u) \supset \mathbb{Z}_p$, a contradiction because

the only subfields of E are \mathbb{Z}_p and E (by Theorem 5). So $x^4 + 1$ is reducible over \mathbb{Z}_p after all. \square

Note that the second half of Example 4 also follows from the corollary to Fermat's theorem (Theorem 8 §1.3).

Before proceeding, recall the following results about a group G :

Lemma 1. Let G be a group and let $x, y \in G$. Then:

- (1) If $o(x) = pq$, $x \in G$ then $o(x^p) = q$.
- (2) If $o(x) = m$, $o(y) = n$, $\gcd(m, n) = 1$ and $xy = yx$, then $o(xy) = mn$.

Proof. (1) is Theorem 5 §2.4. Given x and y as in (2), it is clear that $(xy)^{mn} = 1$. If $(xy)^d = 1$ then $1 = (xy)^{dm} = y^{dm}$. Hence $n \mid dm$, so $n \mid d$ because $\gcd(m, n) = 1$. Similarly $m \mid d$, so $mn \mid d$, again because $\gcd(m, n) = 1$. This proves (2). \blacksquare

Lemma 2. If G is a finite abelian group, let a be an element of G of maximal order. If $o(a) = m$ then $g^m = 1$ for every $g \in G$.

Proof. Suppose $g^m \neq 1$ for some $g \in G$, and write $o(g) = n$. Then n does not divide m so there exists a prime power p^e , $e \geq 1$, such that $p^e \mid n$ but p^e does not divide m (n is a product of prime powers with distinct primes). If we write $n = p^e q$ then $o(g^q) = p^e$ by Lemma 1(1).

On the other hand, write $m = p^t k$ where $t \geq 0$ and p does not divide k . Then $o(a^{p^t}) = k$, again by Lemma 1(1). But k and p^e are relatively prime so Lemma 1(2) gives $o(a^{p^t} g^q) = kp^e > kp^t = m$ because $p^e > p^t$ (p^e does not divide m). This contradiction proves Lemma 2. \blacksquare

We can now prove that the group of units F^* of a finite field F is a cyclic group, a result due to Galois. In fact, we get a stronger result with the same effort.

Theorem 7. Let F be any field. If G is a finite subgroup of the multiplicative group F^* of F , then G is cyclic. In particular, if F is finite, then F^* is cyclic.

Proof. Let $a \in G$ have maximal order m . Then $g^m = 1$ for all $g \in G$ by Lemma 2. Thus, every element of G is a root of $x^m - 1$, so $|G| \leq m$ by Theorem 8 §4.1. But then $m = |\langle a \rangle| \leq |G| \leq m$, so $G = \langle a \rangle$ is cyclic, as required. \blacksquare

Note that Theorem 7 fails if G is infinite, even if G is torsion: consider the unit circle group \mathbb{C}^0 of complex numbers of absolute value 1.

If F is a finite field, a generator for F^* is called a **primitive element** for F . Hence, Theorem 7 asserts that every finite field has a primitive element. In particular, \mathbb{Z}_p has a primitive element for each prime p , called a **primitive root modulo p** . This fact is important in number theory. More generally, if a (possibly infinite) field F has a multiplicative subgroup G of order n , a generator of G (which exists by Theorem 7) is called a **primitive n th root of unity** in F . For example, $e^{2\pi i/n}$ is a primitive n th root of unity in \mathbb{C} for each $n \geq 2$.

If F is a finite field of characteristic p , the existence of a primitive element u in F implies that $F = \mathbb{Z}_p(u)$; in other words, F is a simple extension of \mathbb{Z}_p . We record this fact for future reference.

Corollary 1. $GF(p^n) = \mathbb{Z}_p(u)$, where u is any primitive element for $GF(p^n)$.

Corollary 2. If p is a prime and $n \geq 1$ is an integer, there exists an irreducible polynomial of degree n over \mathbb{Z}_p .

Proof. Write $F = GF(p^n)$ and let $F^* = \langle u \rangle$ by Theorem 7. Here, $F \supseteq \mathbb{Z}_p$, as usual, so let m be the minimal polynomial of u over \mathbb{Z}_p . Then m is irreducible and, as $F = \mathbb{Z}_p(u)$, $\deg m = [\mathbb{Z}_p(u) : \mathbb{Z}_p] = [F : \mathbb{Z}_p] = n$. \blacksquare

Theorem 7 casts new light on the description in Theorem 5 of every subfield K of $F = GF(p^n)$. First, K is uniquely determined by its order because K^* is a subgroup of the cyclic group F^* (and so is unique of its order). Now let u be a primitive element for F so that $F^* = \langle u \rangle$, where $o(u) = p^n - 1$. Because K^* is a subgroup of F^* , it has the form $K^* = \langle u^d \rangle$, where d divides $p^n - 1$. Hence, u^d is a primitive element for K . Moreover, as $|K| = p^m$, where $m|n$ (by Theorem 5), we have $p^m - 1 = |K^*| = |F^*|/d$. Hence,

$$d = \frac{p^n - 1}{p^m - 1} \quad (\text{where } m|n) \quad \text{and} \quad K = \{0\} \cup \langle u^d \rangle.$$

This gives a complete description of the subfields K of F in terms of the divisors m of n and a primitive element u for F .

Exercises 6.4

Throughout these exercises, F denotes a field.

1. Find a primitive element for
 - (a) \mathbb{Z}_{11}
 - (b) \mathbb{Z}_{13}
 - (c) $GF(8)$
 - (d) $GF(9)$
2. Construct a field of order 27 and find a primitive element.
3. Explain why $\mathbb{Z}_2[x]/\langle p \rangle$ and $\mathbb{Z}_2[x]/\langle q \rangle$ are isomorphic if $p = x^3 + x^2 + 1$ and $q = x^3 + x + 1$.
4. If p is a prime, draw the subfield lattice of
 - (a) $GF(p^{12})$
 - (b) $GF(p^{30})$
 - (c) $GF(p^8)$
5. Find a primitive element of $GF(16)$ and use it to write down all the subfields.
6. Find a primitive element of $GF(32)$ and use it to write down all the subfields.
7. Let $E \supseteq F$ be fields. If E is finite, show that $E = F(u)$ for some $u \in E$.
8. Find $[GF(p^n) : GF(p^m)]$, where $m|n$.
9. Describe all the finite subgroups of \mathbb{C}^* .
10. If G and H are subgroups of F^* of order n , show that $G = H$.
11. Show that each element a of $F = GF(p^n)$ has a p th root in F ; that is, $a = b^p$ for some $b \in F$.
12. Let F be a field in which F^* is cyclic. Prove that F is finite.
13. If $E \supseteq \mathbb{Z}_p$ is a field and $u \in E$ is a root of $f \in \mathbb{Z}_p[x]$, show that u^p is also a root.
[Hint: Frobenius automorphism.]
14. Let F be a finite field of characteristic p . If u is a primitive element for F , show that u^p is also a primitive element.
15. Show that $x^2 + x + 1$ is irreducible over $GF(2^n)$ if n is odd. [Hint: If u is a root, compute u^{2^k} for each $k \geq 1$.]
16. Prove (4) of Theorem 2.
17. Let f be a nonconstant polynomial in $F[x]$. Show that f has no repeated root in any splitting field over F if and only if f and f' are relatively prime in $F[x]$.

18. (a) Show that a monic irreducible polynomial f in $F[x]$ has no repeated root in any splitting field over F if and only if $f' \neq 0$ in $F[x]$.
 (b) If $\text{char } F = 0$, show that no irreducible polynomial has a repeated root in any splitting field over F .
19. If $\text{char } F = p$, show that a monic irreducible polynomial f in $F[x]$ has a repeated root in some splitting field if and only if $f = g(x^p)$ for some $g \in F[x]$. [Hint: Exercise 18.]
20. Show that no finite field F is algebraically closed. [Hint: Apply Exercise 17 to $f = x^{q+1} + 1$, where $q = |F|$.]
21. Let p be a prime and write $f = x^p - x - 1$. Show that the splitting field of f over \mathbb{Z}_p is $\mathbb{Z}_p(u)$, where u is any root of f . [Hint: Compute $f(u+a)$, $a \in \mathbb{Z}_p$.]
22. (a) Let f be a monic irreducible polynomial of degree n in $\mathbb{Z}_p[x]$. Show that f divides $x^{p^n} - x$ in $\mathbb{Z}_p[x]$. [Hint: First work over $\mathbb{Z}_p(u)$, $f(u) = 0$. Use the uniqueness in Theorem 4 §4.1.]
 (b) Show that the degree of each monic irreducible divisor f of $x^{p^n} - x$ is a divisor of n . [Hint: Theorem 5.]
 (c) Factor $x^8 - x$ into irreducibles in $\mathbb{Z}_2[x]$.
23. If F is a finite field, show that every element of F is the sum of two squares. [Hint: Given $a \in F$, show that $X = \{u^2 \mid u \in F\}$ and $Y = \{a - u^2 \mid u \in F\}$ each have more than $\frac{1}{2}|F|$ elements.]

6.5 GEOMETRIC CONSTRUCTIONS

Geometry is the only science it hath pleased God to bestow on mankind.

—Thomas Hobbes

The ancient Greeks were good at geometry. However, unlike the analytic geometry of today, which makes extensive use of coordinate systems, the Greeks preferred *synthetic* methods, such as dropping perpendiculars from a point to a line, intersecting lines and curves, and the like. In particular, they were interested in constructions using only compass and straightedge (with no marks on the straightedge). Thus, they allowed drawing lines through two given points, drawing circles with a given center and radius and finding points of intersection of these curves.

For example, the usual method of bisecting an angle uses only these methods. It may come as a surprise that the ancient Greeks were not able to answer the following questions.

- (1) Can any angle be trisected using only compass and straightedge?
- (2) Can any cube be duplicated using only compass and straightedge? That is, can a cube be constructed whose volume is twice that of a given cube?

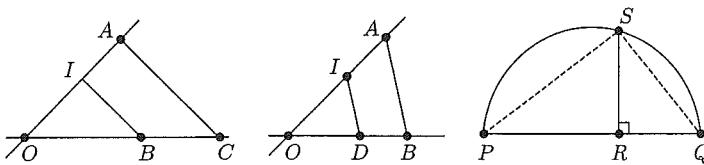
These questions remained unanswered until the nineteenth century, when algebraic methods were applied. The answer to both questions is *no*, as we demonstrate in this section. It is worth noting that, well into the twentieth century, hundreds of people claimed to have solved one of these problems, and some have even gone so far as to publish their “solutions.”

To systematically analyze these questions, the idea of a constructible real number is essential. Suppose that a line segment of finite length is defined to be one

unit in length. Then a real number a is called **constructible** if a line segment of length $|a|$ can be constructed from the unit segment in a finite number of steps using only a compass and straightedge. Note the immediate implication that a number a is constructible if and only if $-a$ is constructible. In fact, we are going to prove that these constructible numbers form a subfield of \mathbb{R} . The essence of the proof is the following Lemma.

Lemma 1. *If $a \geq 0$ and $b \geq 0$ are constructible, then so are $a + b$, $a - b$ (if $a > b$), ab , b/a (if $a \neq 0$), and \sqrt{a} .*

Proof. With the compass, a copy of any finite line segment can be constructed on any given line with any given point as either endpoint. Placing segments end to end shows that $a + b$ is constructible. Similarly, $a - b$ is constructible if $a > b$. □



The diagram on the left-hand side shows the construction for ab (where $a > 1$). Let two nonparallel lines intersect at O and let OI , OA , and OB be segments as shown, with lengths $|OI| = 1$, $|OA| = a$, and $|OB| = b$. Using only compass and straightedge, construct the line through A parallel to IB and let C be the point of intersection of this line and the line through O and B . Then the similarity of the triangles OIB and OAC gives

$$\frac{|OB|}{|OI|} = \frac{|OC|}{|OA|} \quad \text{or} \quad \frac{b}{1} = \frac{|OC|}{a}.$$

Hence, $ab = |OC|$ is constructible. The same argument works if $a < 1$.

The proof that b/a is constructible involves the same setup (middle diagram), except that now the line through I parallel to AB is constructed. Because $a \neq 0$, this line meets the line through O and B at D , say. Then the similarity of OID and OAB gives $|OD| = b/a$, so b/a is constructible.

Finally, to show that \sqrt{a} is constructible, consider a semicircle with diameter PQ of length $a + 1$ (right-hand side diagram) and let R be the point of PQ such that $|PR| = a$. Using a compass and straightedge, construct the line through R perpendicular to the diameter and let this line meet the arc of the circle at S . It is a theorem of geometry that the angle PSQ is a right angle, so the triangles PSR and SQR are similar. Hence, $|SR|/|PR| = |RQ|/|SR|$, that is, $|SR|^2 = |PR||RQ| = a \cdot 1 = a$. Thus, $\sqrt{a} = |SR|$ is constructible. □

Although Lemma 1 deals only with nonnegative constructible numbers, the reader can now easily supply the proof of Theorem 1.

Theorem 1. *The set of all constructible numbers is a subfield of \mathbb{R} .*

Note that every rational number is constructible.

With Theorem 1 in hand, we attack the Greek construction problems as follows. We begin by showing that the minimal polynomial over \mathbb{Q} of every constructible number has degree a power of 2. Then we prove that a given construction is

impossible by showing that it would allow the construction of a number with minimal polynomial having degree not a power of 2. As we shall see, this latter step is quite easy for the two Greek questions mentioned earlier, so we turn to the algebraic condition on the constructible numbers.

Let C denote the field of constructible numbers. If $a \in C$, then a is the distance between two points in the plane that have been obtained by a finite series of compass and straightedge constructions, beginning with points with rational coordinates. Hence, we consider an arbitrary subfield F of C and investigate the nature of the points obtained by a single compass and straightedge construction, beginning with points whose coordinates lie in F (called F -points for short). The straightedge provides lines through pairs of F -points and the compass provides circles centered at F -points with radius in F (called F -lines and F -circles, respectively). The only way to construct new points is as points of intersection of two F -lines, of an F -line and an F -circle, or of two F -circles. We can easily verify (Exercise 2) that the equations of F -lines and F -circles have the following form:

$$\begin{aligned} F\text{-lines: } & ax + by = c, & a, b, \text{ and } c \text{ in } F, \\ F\text{-circles: } & x^2 + y^2 + ax + by = c, & a, b, \text{ and } c \text{ in } F. \end{aligned}$$

It follows that if two F -lines intersect, the point of intersection is an F -point (Exercise 2). However, finding the intersection points (x, y) of an F -line and an F -circle (if they exist) leads to a quadratic equation for x or y with coefficients in F . Hence, by the quadratic formula, x and y lie in an extension $F(\sqrt{a})$ of F , where $a \in F$ and $a > 0$ (Exercise 2). Finally, the intersection points (if any) of two F -circles can be obtained as the intersections of one of the circles with an F -line (the one through the points in question). Hence, in all cases, a compass and straightedge construction beginning with F -points leads to $F(\sqrt{a})$ -points, where $a \in F$, $a > 0$. Observe that \sqrt{a} is constructible by Lemma 1, so $F(\sqrt{a}) \subseteq C$ by Theorem 1. Finally, note that

$$[F(\sqrt{a}) : F] = 2 \text{ or } 1,$$

depending on whether $x^2 - a$ is irreducible or not in $F[x]$.

Now suppose that a is any constructible number. Then $|a|$ can be constructed as the distance between two C -points P and Q , where C is the field of constructible numbers and where these points are obtained by a series of compass and straightedge constructions beginning with \mathbb{Q} -points. By the preceding discussion, the first of these constructions produces F_1 -points, where F_1 is a field, $C \supseteq F_1 \supseteq \mathbb{Q}$, and $[F_1 : \mathbb{Q}] = 1$ or 2. Then the second construction yields F_2 -points, where F_2 is a field, $C \supseteq F_2 \supseteq F_1$, and $[F_2 : F_1] = 1$ or 2. The process continues to create a chain of fields $\mathbb{Q} = F_0 \subseteq F_1 \subseteq F_2 \subseteq \cdots \subseteq C$. Suppose that $m - 1$ constructions are needed to obtain P and Q , so that P and Q are F_{m-1} -points. Because a is the distance between P and Q , the distance formula shows that $a \in F_{m-1}(\sqrt{b})$, where $b \in F_{m-1}$ and $b > 0$. Writing $F_m = F_{m-1}(\sqrt{b})$, this means that $[F_m : F_{m-1}] = 1$ or 2. Hence, we have constructed a finite chain

$$\mathbb{Q} = F_0 \subseteq F_1 \subseteq F_2 \subseteq \cdots \subseteq F_{m-1} \subseteq F_m$$

of fields where $[F_k : F_{k-1}] = 1$ or 2 for each k and where $a \in F_m$. Now the multiplication theorem (Theorem 5 §6.2) gives

$$[F_m : \mathbb{Q}] = [F_m : F_{m-1}] \cdots [F_2 : F_1][F_1 : F_0],$$

so, as each $[F_k : F_{k-1}] = 1$ or 2 , $[F_m : \mathbb{Q}]$ is a power of 2 . But $a \in F_m$ implies that $\mathbb{Q}(a) \subseteq F_m$, so the multiplication theorem implies that $[\mathbb{Q}(a) : \mathbb{Q}]$ is a power of 2 (being a divisor of $[F_m : \mathbb{Q}]$). Because $[\mathbb{Q}(a) : \mathbb{Q}]$ is the degree of the minimal polynomial of a over \mathbb{Q} (Theorem 4 §6.2), this condition proves

Theorem 2. *If a is a constructible number, then $[\mathbb{Q}(a) : \mathbb{Q}] = 2^k$ for some $k \geq 0$. In particular, the minimal polynomial of a over \mathbb{Q} has degree 2^k .*

Theorem 2 implies that every constructible number is algebraic over \mathbb{Q} . Moreover, the argument leading to Theorem 2 actually provides a characterization of the constructible numbers: A real number a is constructible if and only if a chain $\mathbb{Q} = F_0 \subseteq F_1 \subseteq F_2 \subseteq \dots \subseteq F_m$ of subfields of \mathbb{R} exists such that $a \in F_m$ and $[F_k : F_{k-1}] = 1$ or 2 for each k (Exercise 8).

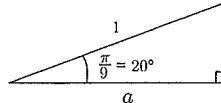
Theorem 2 provides the means to easily settle the classical construction questions posed at the beginning of this section.

Corollary 1. *It is impossible to duplicate a cube of side 1 using only a compass and straightedge.*

Proof. If it were possible, then a cube of volume 2 could be constructed. A side of this cube has length $\sqrt[3]{2}$, which would mean that $\sqrt[3]{2}$ is constructible. But $x^3 - 2$ is irreducible over \mathbb{Q} (by the Eisenstein Criterion, Theorem 8 §4.2), so it is the minimal polynomial of $\sqrt[3]{2}$. Because the degree 3 is not a power of 2, Theorem 2 shows that $\sqrt[3]{2}$ cannot be constructed. ■

Corollary 2. *It is impossible to trisect $\pi/3$ by compass and straightedge.⁷⁴*

Proof. Write $\alpha = \cos(\pi/9)$. If trisection of $\pi/3$ were possible, the right triangle in the figure (with hypotenuse of length 1) could be constructed, and hence a would be a constructible number. We show that this is not so by proving



that the degree of a over \mathbb{Q} is 3. To accomplish this, recall the trigonometric identity $\cos 3\theta = 4\cos^3 \theta - 3\cos \theta$ (see Exercise 13, Appendix A). If we take $\theta = \pi/9$, this becomes $\frac{1}{2} = 4a^3 - 3a$. Hence, a is a root of $m = 8x^3 - 6x - 1$, which is irreducible over \mathbb{Q} by Theorem 1 §4.2 because it has no root (by Theorem 9 §4.1). Thus, $\frac{1}{8}m$ is the minimal polynomial of a over \mathbb{Q} , and so the degree is 3 as asserted. ■

Another famous problem of the ancient Greeks is whether it is possible using only compass and straightedge to square a circle—that is, to construct a square with area equal to that of a given circle. This too is ruled out by Theorem 2, together with a result of Ferdinand von Lindemann.

Corollary 3. *It is impossible with a compass and straightedge to construct a square with area equal to the area of a circle of radius 1.*

Proof. The area of such a square would be π , so the length $\sqrt{\pi}$ of a side of this square would be constructible. In particular, $\sqrt{\pi}$ would be algebraic over \mathbb{Q} , so π would be algebraic. But π is transcendental over \mathbb{Q} by a famous theorem of Lindemann. ■

⁷⁴Archimedes showed that if we are allowed to *mark* the straightedge, it is possible to trisect any angle.

Exercises 6.5

1. Give a compass and straightedge construction for each of
 - (a) A line parallel to a given line through a given point.
 - (b) A line perpendicular to a given line through a given point.

(These are used in the proof of Lemma 1.)
2. Let F be a subfield of the field of constructible numbers. Show that
 - (a) Each F -line has equation $ax + by = c$, where $a, b, c \in F$.
 - (b) Each F -circle has equation $x^2 + y^2 + ax + by = c$, where $a, b, c \in F$.
 - (c) The intersection (if any) of two F -lines is an F -point.
 - (d) The intersections (if any) of an F -line and an F -circle are $F(\sqrt{a})$ -points, where $a \in F$ and $a > 0$.
3. Can an angle of $\pi/4 = 45^\circ$ be trisected using only a compass and straightedge? Support your answer.
4. Can an angle of 40° be constructed? Support your answer.
5. Can a sphere be cubed? That is, can a cube be constructed whose volume equals that of a given sphere? Support your answer.
6. Can a cube be tripled? That is, can a cube be constructed whose volume is three times that of a given cube? Support your answer.
7. (a) Show that $\sin \theta$ is constructible if and only if $\cos \theta$ is constructible.
 (b) Show that $\cos 2\theta$ is constructible if and only if $\cos \theta$ is constructible.
8. Show that a real number a is constructible if and only if a finite chain of subfields $\mathbb{Q} = F_0 \subseteq F_1 \subseteq \dots \subseteq F_m$ of \mathbb{R} exists with $a \in F_m$ and $[F_k : F_{k-1}] = 1, 2$ for each k .
9. Show that a regular heptagon (seven-sided polygon with vertices equally spaced on a circle) is not constructible with a compass and straightedge. [Hint: $64x^7 - 112x^5 + 56x^3 - 7x - 1 = (8x^3 + 4x^2 - 4x - 1)(8x^4 - 4x^3 - 8x^2 + 3x + 1)$.]

6.6 THE FUNDAMENTAL THEOREM OF ALGEBRA⁷⁵

The fundamental theorem of algebra is the assertion that the field \mathbb{C} of complex numbers is algebraically closed. This result was first proved by Gauss in his Ph.D. dissertation, and many proofs of this result are now known. However, no proof is entirely algebraic in nature; that is, each proof involves some analytic property of polynomials. The proof we give uses only one nonalgebraic fact:

If a polynomial f in $\mathbb{R}[x]$ has odd degree, then f has a real root.

This fact, known to every calculus student, depends on the continuity of f when regarded as a function $\mathbb{R} \rightarrow \mathbb{R}$. Because f has odd degree, there are real numbers a and b such that $f(a) > 0$ and $f(b) < 0$. The graph of f is a continuous curve, so it must cut the x -axis at some value u between a and b . Hence $f(u) = 0$, and u is the desired real root.

The algebraic prerequisites for our proof are the existence of splitting fields and a result about symmetric polynomials. A polynomial $f(x_1, \dots, x_n)$ in n variables

⁷⁵This section requires results from Section 4.5 on symmetric polynomials.

is called **symmetric** if it is unchanged when the variables are permuted; that is,

$$f(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = f(x_1, x_2, \dots, x_n) \text{ for all } \sigma \in S_n.$$

Thus, $f(x_1, x_2) = x_1^2 + x_2^2$ is symmetric, as is $f(x_1, x_2, x_3) = x_1 x_2 x_3$. If $1 \leq k \leq n$, the **elementary symmetric polynomial** $s_k = s_k(x_1, \dots, x_n)$ is defined to be the sum of all possible products of k of the variables x_1, \dots, x_n . More formally,

$$s_k = s_k(x_1, \dots, x_n) = \sum_{i_1 < i_2 < \dots < i_k} x_{i_1} x_{i_2} \cdots x_{i_k}.$$

We define $s_0(x_1, \dots, x_n) = 1$. Hence, for example,

$$\begin{aligned} s_1(x_1, x_2, \dots, x_n) &= x_1 + x_2 + \cdots + x_n, \\ s_n(x_1, x_2, \dots, x_n) &= x_1 x_2 \cdots x_n, \\ s_2(x_1, x_2, x_3, x_4) &= x_1 x_2 + x_1 x_3 + x_1 x_4 + x_2 x_3 + x_2 x_4 + x_3 x_4, \\ s_3(x_1, x_2, x_3, x_4) &= x_1 x_2 x_3 + x_1 x_2 x_4 + x_1 x_3 x_4 + x_2 x_3 x_4. \end{aligned}$$

The importance of the polynomials $s_k = s_k(x_1, x_2, \dots, x_n)$ for the splitting of polynomials lies in the fact that

$$\begin{aligned} (x - x_1)(x - x_2) \cdots (x - x_n) \\ = x^n - s_1 x^{n-1} + s_2 x^{n-2} + \cdots + (-1)^{n-1} s_{n-1} x + (-1)^n s_n. \end{aligned}$$

If $f(x_1, \dots, x_n)$ is any polynomial, it is clear that $f(s_1, s_2, \dots, s_n)$ is symmetric. The remarkable thing is that the converse holds (Theorem 4 §4.5).

Theorem 1. Fundamental Theorem on Symmetric Polynomials. Every symmetric polynomial over a ring R is a polynomial $f(s_1, s_2, \dots, s_n)$ over R in the elementary symmetric polynomials s_1, \dots, s_n . In fact, this representation is unique.

For example, if $n = 3$, the symmetric polynomial $x_1^2 + x_2^2 + x_3^2$ has the representation

$$x_1^2 + x_2^2 + x_3^2 = (x_1 + x_2 + x_3)^2 - 2(x_1 x_2 + x_1 x_3 + x_2 x_3) = s_1^2 - 2s_2.$$

Moreover, this result holds for any number of variables as the reader can verify. Other examples are given in Section 4.5.

Theorem 2. Fundamental Theorem of Algebra. The field \mathbb{C} of complex numbers is algebraically closed.

Proof. We show that a nonconstant polynomial f in $\mathbb{C}[x]$ has a root in \mathbb{C} . First, we show that it suffices to prove this property for real polynomials. If \bar{f} is obtained from f by conjugating every coefficient, then $g = f \bar{f}$ has real coefficients, as is easily verified, and if $g(u) = 0$, $u \in \mathbb{C}$, then either $f(u) = 0$ or $0 = \bar{f}(u) = f(\bar{u})$. Hence, u or \bar{u} is a root of f .

So let f be a nonconstant polynomial in $\mathbb{R}[x]$ and write $\deg f = d = 2^n m$, where m is odd. We show that f has a root in \mathbb{C} by induction on $n \geq 0$. If $n = 0$, then f has odd degree and so has a root in \mathbb{R} . If $n \geq 1$, regard f as an element in $\mathbb{C}[x]$ and let $E \supseteq \mathbb{C}$ be a splitting field for f . Hence,

$$f = a(x - u_1)(x - u_2) \cdots (x - u_d), \quad \text{where } a \in \mathbb{R} \text{ and each } u_i \in E.$$

It suffices to show that $u_i \in \mathbb{C}$ for some i . We have

$$f = a x^d - a s_1(u_1, \dots, u_d) x^{d-1} + a s_2(u_1, \dots, u_d) x^{d-2} + \cdots + (-1)^d a s_d(u_1, \dots, u_d).$$

Hence, $s_k(u_1, \dots, u_d) \in \mathbb{R}$ for each $k = 1, 2, \dots, d$ because $f \in \mathbb{R}[x]$.

Given $1 \leq h \in \mathbb{Z}$, consider the following polynomial in $\mathbb{R}[x][x_1, \dots, x_d]$:

$$f_h(x; x_1, \dots, x_d) = \prod_{1 \leq i < j \leq d} (x - x_i - x_j - hx_i x_j). \quad (*)$$

For fixed x , $(*)$ is a symmetric polynomial in the variables x_1, x_2, \dots, x_d and so, by Theorem 1, it is a polynomial in $s_1(x_1, \dots, x_d), \dots, s_d(x_1, \dots, x_d)$ with coefficients in $\mathbb{R}[x]$. Because $s_1(u_1, \dots, u_d), \dots, s_d(u_1, \dots, u_d)$ are in \mathbb{R} , this means that the polynomial $f_h(x) = f_h(x; u_1, \dots, u_d)$ is in $\mathbb{R}[x]$. Moreover,

$$\deg(f_h) = \binom{d}{2} = \frac{1}{2}d(d-1) = 2^{n-1}m(2^n m - 1),$$

so, by induction, f_h has a root in \mathbb{C} for each $h \geq 1$. Hence, $(*)$ implies that, given $h \geq 1$, $u_i + u_j + hu_i u_j \in \mathbb{C}$ for some i and j , with $1 \leq i < j \leq d$. As the number of such pairs (i, j) is finite, integers $h \neq h'$ exist such that both $u_i + u_j + hu_i u_j$ and $u_i + u_j + h'u_i u_j$ lie in \mathbb{C} . Then both $u_i + u_j$ and $u_i u_j$ are in \mathbb{C} , so $(x - u_i)(x - u_j) \in \mathbb{C}[x]$. But this polynomial splits in \mathbb{C} by the quadratic formula, so u_i and u_j are in \mathbb{C} . Because they are roots of f , the proof is complete. ■

A closer scrutiny of the proof of Theorem 2 reveals that we have proved slightly more: Let $C \supseteq F$ be fields and assume that $C = F(i)$, where $i^2 + 1 = 0$, and that

- (1) F has characteristic 0.
- (2) Each element of C has a square root in C .
- (3) Each polynomial in $F[x]$ of odd degree has a root in C .

Then C is algebraically closed.

6.7 AN APPLICATION TO CYCLIC AND BCH CODES

There is no branch of mathematics, however abstract, which may not someday be applied to phenomena of the real world.

—Nicolai Ivanovich Lobachevski

We introduced coding theory in Section 2.11, where we discussed binary linear codes. Recall that the direct product of n copies of \mathbb{Z}_2 is denoted B^n and that elements of B^n are called words and written as strings of 0's and 1's (called bits). Thus, $B^2 = \{00, 01, 10, 11\}$. In general, B^n is an additive group of order 2^n . A subgroup $C \subseteq B^n$ of order 2^k , $k \neq 0$, is called a binary linear code or an (n, k) -code for short. In this section, we discuss an important class of linear codes, called cyclic codes. These codes are useful because they can be implemented by a simple electronic circuit called a feedback shift register (discussion of which is beyond the scope of this book). Our interest in these codes is twofold: (1) their analysis provides an application of the theory of rings, polynomials, and fields and (2) they include the so-called BCH codes—one of the most widely used classes of error-correcting codes.

If $C \subseteq B^n$ is a code, a word in C is denoted $a_0 a_1 a_2 \cdots a_{n-1}$, where the bits a_i are in $\mathbb{Z}_2 = \{0, 1\}$. The reason for this choice of subscripts will soon be apparent.

A code is called **cyclic** if it is closed under **cyclic shifts**; that is, if $a_0a_1a_2 \cdots a_{n-1}$ is in C , then $a_{n-1}a_0a_1a_2 \cdots a_{n-2}$ is also in C .

Example 1. $\{000, 111\}$ and $\{000, 110, 011, 101\}$ are cyclic by inspection.

Example 2. The set of words in B^n of even parity (even number of 1-bits) is a cyclic code of order 2^{n-1} . For $n = 4$, it is

$$C = \{0000, 1100, 0110, 0011, 1010, 0101, 1001, 1111\}.$$

The theory of rings enters the picture as an elegant means of describing these cyclic codes. Let x be an indeterminate over \mathbb{Z}_2 and consider the principal ideal $\langle 1 - x^n \rangle$ of all multiples $1 - x^n$ in the polynomial ring $\mathbb{Z}_2[x]$. The factor ring is denoted

$$B_n = \frac{\mathbb{Z}_2[x]}{\langle 1 - x^n \rangle}.$$

Recall (Theorem 2 §4.3) that the ring B_n can be described as

$$B_n = \{a_0 + a_1t + \cdots + a_{n-1}t^{n-1} \mid a_i \in \mathbb{Z}_2, t^n = 1\}.$$

The operations in B_n are the same as for polynomials, except that $t^n = 1$. The map

$$\theta : \mathbb{Z}_2[x] \rightarrow B_n \quad \text{given by} \quad \theta(f) = f(t)$$

is an onto ring homomorphism with $\theta(x) = t$ and $\ker \theta = \langle 1 - x^n \rangle$. Hence,

$$B_n = \{f(t) \mid f \in \mathbb{Z}_2[x]\}.$$

Moreover, $\{1, t, \dots, t^{n-1}\}$ is a basis of B_n as a vector space over the field \mathbb{Z}_2 , so the additive groups B_n and B^n are isomorphic via the correspondence

$$a_0 + a_1t + \cdots + a_{n-1}t^{n-1} \leftrightarrow a_0a_1 \cdots a_{n-1}.$$

For example, if $n = 5$, some typical correspondences are

$$\begin{aligned} 1 + t^2 + t^3 &\text{ corresponds to } 10110, \\ 1 &\text{ corresponds to } 10000, \\ 1 + t + t^2 + t^3 + t^4 &\text{ corresponds to } 11111. \end{aligned}$$

Because of this isomorphism, we think of codes as additive subgroups of B^n or B_n . We call these the **word form** and the **polynomial form** of the code, respectively, and use both points of view in this section. The word form of a code is useful when matrix multiplication is used for encoding (see Section 2.11). However, the polynomial form of a code has the advantage that the extra ring structure of B_n is useful for describing cyclic codes.

Indeed, if $C \subseteq B_n$ is a code and $f(t) = a_0 + a_1t + \cdots + a_{n-1}t^{n-1}$ is an element in C , the cyclic shift of $f(t)$ is

$$a_{n-1} + a_0t + a_1t^2 + \cdots + a_{n-2}t^{n-2} = tf(t)$$

using the multiplication in B_n and the fact that $t^n = 1$. Hence,

$$C \text{ is cyclic} \quad \text{if and only if} \quad tC \subseteq C,$$

where, of course, $tC = \{tf(t) \mid f(t) \in C\}$. But because C is an additive subgroup of B_n , the condition $tC \subseteq C$ means C is an ideal of the (commutative) ring B_n .

This is wonderful news because the ideals of B_n are easy to describe. Recall the onto ring homomorphism $\theta : \mathbb{Z}_2[x] \rightarrow B_n$ given by $\theta(f) = f(t)$ for $f \in \mathbb{Z}_2[x]$, with $\ker \theta = \langle 1 - x^n \rangle$. If C is an ideal of B_n define

$$A = \{f \in \mathbb{Z}_2[x] \mid f(t) \in C\}.$$

It is routine to verify that A is an ideal of $\mathbb{Z}_2[x]$ such that $\ker \theta \subseteq A$ and that $C = \theta(A)$. But A is a principal ideal because \mathbb{Z}_2 is a field. More precisely, Theorem 1 §4.3 shows that if g is a nonzero polynomial in A of minimal degree, then g is uniquely determined by A (it is automatically monic because the field is \mathbb{Z}_2), and

$$A = \langle g \rangle = \{qg \mid q \in \mathbb{Z}_2[x]\} = \mathbb{Z}_2[x]g.$$

Moreover, $\ker \theta \subseteq A$ means that $\langle 1 - x^n \rangle \subseteq \langle g \rangle$, so g is a divisor of $1 - x^n$ in $\mathbb{Z}_2[x]$. Hence, every ideal C of B_n has the form

$$C = \theta(A) = \langle g(t) \rangle = \{q(t)g(t) \mid q(t) \in B_n\} = B_ng(t),$$

where g divides $1 - x^n$. Theorem 1 summarizes this discussion.

Theorem 1. *The following conditions are equivalent for a code $C \subseteq B_n$:*

- (1) C is cyclic.
- (2) $tC \subseteq C$.
- (3) C is an ideal of the ring B_n .

In this case, a divisor g of $1 - x^n$ exists in $\mathbb{Z}_2[x]$ such that

$$C = \langle g(t) \rangle = \{q(t)g(t) \mid q(t) \text{ in } B_n\} = \{f(t) \mid g \text{ divides } f \text{ in } \mathbb{Z}_2[x]\}.$$

Moreover, g is the unique polynomial in $\mathbb{Z}_2[x]$ of lowest degree such that $g(t) \in C$. Finally, if $\langle f(t) \rangle$ is another such code, where f divides $1 - x^n$, then $\langle g(t) \rangle \subseteq \langle f(t) \rangle$ if and only if f divides g in $\mathbb{Z}_2[x]$.

Proof. Only the last statement remains to be proved. Suppose that $\langle g(t) \rangle \subseteq \langle f(t) \rangle$ so that $g(t) = q(t)f(t)$ in B_n . Then $g - qf$ lies in $\ker \theta = \langle 1 - x^n \rangle$. Since f divides $1 - x^n$, this implies that f divides g . The converse is clear. ■

To illustrate Theorem 1, consider again the code

$$C = \{0, 1 + t, t + t^2, t^2 + t^3, 1 + t^2, t + t^3, 1 + t^3, 1 + t + t^2 + t^3\}$$

in Example 2. It is generated by $1 + t$ because $g = 1 + x$ is the nonzero polynomial of least degree such that $g(t)$ is in C . In general, a cyclic code can have more than one generator (both $t + t^2$ and $1 + t^3$ generate C), but there is only one of least degree. This unique polynomial is called the **minimal generator** of C .

Hence, determining the cyclic codes in B_n comes down to identifying all the divisors of $1 - x^n = 1 + x^n$ in $\mathbb{Z}_2[x]$. These divisors, in turn, are determined by the

factorization of $1 + x^n$ into irreducible factors in $\mathbb{Z}_2[x]$. The factorizations for the first few values of n are

$$\begin{aligned}1 + x^2 &= (1 + x)^2, \\1 + x^3 &= (1 + x)(1 + x + x^2), \\1 + x^4 &= (1 + x)^4, \\1 + x^5 &= (1 + x)(1 + x + x^2 + x^3 + x^4), \\1 + x^6 &= (1 + x)^2(1 + x + x^2)^2, \\1 + x^7 &= (1 + x)(1 + x + x^3)(1 + x^2 + x^3).\end{aligned}$$

Recall that quadratic and cubic polynomials are irreducible if they have no root in \mathbb{Z}_2 , but no such simple test exists if the degree is greater than 3.

We note in passing that if F is a field, finding the irreducible factors of $1 - x^n$ in $F[x]$ begins by factoring it as a product of cyclotomic polynomials. We do so for $F = \mathbb{Q}$ in Section 10.4, where we show that the cyclotomic polynomials themselves are irreducible. However, if $F = \mathbb{Z}_2$, each cyclotomic polynomial factors into irreducible polynomials of the same degree. Discussion of this topic is beyond the scope of this book.⁷⁶

Example 3. Recalling that $2 = 0$ in \mathbb{Z}_2 , we get $1 + x^4 = (1 + x)^4$ in $\mathbb{Z}_2[x]$. Hence, the divisors of $1 + x^4$ are $1, 1 + x, (1 + x)^2$, and $(1 + x)^3$. Because $(1 + x)^2 = 1 + x^2$ and $(1 + x)^3 = 1 + x + x^2 + x^3$, the cyclic codes in B_4 are

$$\begin{aligned}\langle 1 \rangle &= B_4, \\\langle 1 + t \rangle &= \{0, 1 + t, t + t^2, t^2 + t^3, 1 + t^3, 1 + t^2, t + t^3, 1 + t + t^2 + t^3\}, \\\langle 1 + t^2 \rangle &= \{0, 1 + t^2, t + t^3, 1 + t + t^2 + t^3\}, \\\langle 1 + t + t^2 + t^3 \rangle &= \{0, 1 + t + t^2 + t^3\}.\end{aligned}$$

Note that $\langle 1 + t \rangle$ corresponds to the code in Example 2.

Example 4. For any n , $1 - x^n = 1 + x^n = (1 + x)(1 + x + \dots + x^{n-1})$ in $\mathbb{Z}_2[x]$. Hence, there are always two cyclic codes:

- $\langle 1 + t \rangle$ is the ideal of polynomials of even parity (coefficients sum to 0).
- $\langle 1 + t + \dots + t^{n-1} \rangle = \{0, 1 + t + \dots + t^{n-1}\}$.

Example 5. We have $1 - x^5 = 1 + x^5 = (1 + x)(1 + x + x^2 + x^3 + x^4)$, and $1 + x + x^2 + x^3 + x^4$ is irreducible (Exercise 10). Hence, B_5 has three cyclic codes:

$$\begin{aligned}\langle 1 \rangle &= B_5, \\\langle 1 + t \rangle &= \text{the polynomials of even parity,} \\\langle 1 + t + t^2 + t^3 + t^4 \rangle &= \{0, 1 + t + t^2 + t^3 + t^4\}.\end{aligned}$$

A code $C \subseteq B_n$ is a \mathbb{Z}_2 -subspace of B_n , so we may speak of the dimension of C over \mathbb{Z}_2 . We write it as $\dim_{\mathbb{Z}_2} C$ or simply $\dim C$ when no confusion can result. Then

$$|C| = 2^k, \quad \text{where} \quad \dim C = k.$$

We now let $C = \langle g(t) \rangle$ be a cyclic code, where g is a divisor of $1 - x^n = 1 + x^n$. If $m = \deg g$, we are going to show that $\dim C = n - m$ and hence that $|C| = 2^{n-m}$.

⁷⁶See, for example Lidl, R. and Neiderreiter, H., *Introduction to Finite Fields and Their Applications*, Cambridge, England: Cambridge University Press, 1986, Section 2.4.

Recall some ring theoretic notation: If R is a commutative ring and $a \in R$, the set

$$\text{ann } a = \{r \in R \mid ra = 0\} \quad \text{is an ideal of } R$$

(called the **annihilator** of a in R). These ideals play a basic role in B_n .

Lemma 1. Let g be a divisor of $1 - x^n$ in $\mathbb{Z}_2[x]$, say $1 - x^n = gh$. Then, in the ring B_n , $\text{ann } g(t) = \langle h(t) \rangle$.

Proof. We have $h(t)g(t) = 1 - t^n = 0$, so $h(t) \in \text{ann } g(t)$. Hence, $\langle h(t) \rangle \subseteq \text{ann } g(t)$. Conversely, if $f(t)g(t) = 0$, then $fg \in \ker \theta = \langle 1 - x^n \rangle$, say $fg = q(1 - x^n)$. As $1 - x^n = hg$, it follows that $f = qh$. Hence, $f(t) \in \langle h(t) \rangle$, as required. ■

Theorem 2. Let $C = \langle g(t) \rangle$ be a cyclic code in B_n , where g divides $1 - x^n$ in $\mathbb{Z}_2[x]$, and write $m = \deg g$ and $k = n - m$. Then

$$X = \{g(t), tg(t), t^2g(t), \dots, t^{k-1}g(t)\}$$

is a \mathbb{Z}_2 -basis of C . In particular, $|C| = 2^k = 2^{n-m}$, so C is an (n, k) -code.

Proof. It suffices to prove that X is a \mathbb{Z}_2 -basis of C . Write $1 - x^n = hg$ so that $\deg h = n - m = k$. Given an element $f(t)$ of C , say $f(t) = q(t)g(t)$, write $q = ph + r$ in $\mathbb{Z}_2[x]$, where $r = a_0 + a_1x + \dots + a_{k-1}x^{k-1}$, $a_i \in \mathbb{Z}_2$. Then $h(t)g(t) = 0$ by Lemma 1, so $f(t) = r(t)g(t)$. Hence, X is a spanning set for C . To see that X is linearly independent, suppose that

$$a_0g(t) + a_1tg(t) + \dots + a_{k-1}t^{k-1}g(t) = 0, \quad a_i \in \mathbb{Z}_2.$$

Write $f = a_0 + a_1x + \dots + a_{k-1}x^{k-1}$; it suffices to show that $f = 0$ in $\mathbb{Z}_2[x]$. But $f(t)g(t) = 0$, so $f(t) \in \langle h(t) \rangle$ by Lemma 1. Hence, h divides f in $\mathbb{Z}_2[x]$ by Theorem 1, which means that $f = 0$ (otherwise, $k = \deg h \leq \deg f \leq k - 1$). ■

Matrix Description

The linear codes in Section 2.11 were described using binary matrices. As already noted, B^n is an n -dimensional vector space over \mathbb{Z}_2 , and an (n, k) -code $C \subseteq B^n$ is nothing but a k -dimensional subspace. If $\{w_0, w_1, \dots, w_{k-1}\}$ is a basis of C , let

$$G = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{k-1} \end{bmatrix}$$

be the $k \times n$ matrix whose rows are the words w_i . For $u = a_0a_1 \dots a_{k-1}$ in B^k ,

$$uG = [a_0 \quad a_1 \quad \dots \quad a_{k-1}] \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{k-1} \end{bmatrix} = a_0w_0 + a_1w_1 + \dots + a_{k-1}w_{k-1},$$

so, as C is spanned by the words w_0, w_1, \dots, w_{k-1} , we have

$$C = \{uG \mid u \in B^k\}.$$

Hence, as in Section 2.11, G is called a **generator matrix**⁷⁷ for the code C . Similarly, if H is an $n \times (n - k)$ matrix such that

$$C = \{w \in B^n \mid wH = 0\},$$

then H is called a **parity check matrix** for the code C . Both methods of describing C are useful, and we can easily find such matrices if C is a cyclic code.

In fact, let $C = \langle g(t) \rangle$ be a cyclic code in B_n , where $1 - x^n = gh$ in $\mathbb{Z}_2[x]$. Write $\deg g = m$, $\deg h = k$, so that $n = m + k$. Then $g(t)$ and $h(t)$ give rise to a generator matrix G and a parity check matrix H for (the word form of) C . If $g(t) = g_0 + g_1t + \cdots + g_mt^m$, we define a $k \times n$ binary matrix

$$G = \begin{bmatrix} g(t) \\ tg(t) \\ \vdots \\ t^{k-1}g(t) \end{bmatrix} = \begin{bmatrix} g_0 & g_1 & g_2 & \cdots & g_m & 0 & \cdots & 0 \\ 0 & g_0 & g_1 & \cdots & g_{m-1} & g_m & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & g_0 & g_1 & \cdots & \cdots & g_m \end{bmatrix}, \quad (*)$$

where the rows of G are the first k cyclic shifts of the coefficients of $g(t)$. Given a word $u = a_0a_1 \cdots a_{k-1}$ in B^k , and being somewhat facile with the notation, we get

$$uG = [a_0 \ a_1 \ \cdots \ a_{k-1}] \begin{bmatrix} g(t) \\ tg(t) \\ \vdots \\ t^{k-1}g(t) \end{bmatrix} = (a_0 + a_1t + \cdots + a_{k-1}t^{k-1})g(t).$$

Hence, G is a generator matrix for (the word form of) $C = \langle g(t) \rangle$.

Lemma 1 (with g and h interchanged) shows that $C = \{f(t) \mid f(t)h(t) = 0\}$. Hence, not surprisingly, a parity check matrix for C comes from $h(t)$ in a similar way. If $h(t) = h_0 + h_1t + \cdots + h_kt^k$, define the $n \times m$ matrix

$$H = \begin{bmatrix} 0 & 0 & \cdots & h_k \\ \vdots & \vdots & \ddots & \vdots \\ 0 & h_k & & \\ h_k & h_{k-1} & h_1 & \\ \vdots & \vdots & \ddots & h_0 \\ h_2 & h_1 & \ddots & 0 \\ h_1 & h_0 & \ddots & \vdots \\ h_0 & 0 & \cdots & 0 \end{bmatrix}, \quad (**)$$

where the columns are the first m cyclic shifts (bottom up) of the coefficients of $h(t)$. The proof that H is a parity check matrix depends on Lemma 2.

⁷⁷We do not insist that the first k columns of G form the $k \times k$ identity matrix; that is, we do not insist that G is a *standard* generator matrix for C (see the definition preceding Theorem 6 §2.11). This restriction is not severe (Exercise 19).

Lemma 2. If G and H are as in (*) and (**), then $GH = 0$.

Proof. As might be expected, the reason is that $g(t)h(t) = 0$ in B_n . Write

$$g(t) = \sum_{i=0}^{n-1} g_i t^i \quad \text{and} \quad h(t) = \sum_{i=0}^{n-1} h_i t^i,$$

where $g_{m+1} = \dots = g_{n-1} = 0$ and $h_{k+1} = \dots = h_{n-1} = 0$. As $t^n = 1$, the coefficient of t^p in the product $g(t)h(t) = 0$ is

$$g_0 h_p + g_1 h_{p-1} + \dots + g_p h_0 + g_{p+1} h_{n-1} + \dots + g_{n-1} h_{p+1} = 0.$$

Taking subscripts modulo n , this expression can be compactly written as

$$\sum_{i+j=p} g_i h_j = 0 \quad \text{for all } p = 0, 1, \dots, n-1.$$

Now the matrix product of a typical row of G and a typical column of H is

$$\begin{bmatrix} g_p & g_{p+1} & \cdots & g_{p+n-1} \end{bmatrix} \begin{bmatrix} h_{q+n-1} \\ \vdots \\ h_{q+1} \\ h_q \end{bmatrix} = \sum_{m=0}^{n-1} g_{p+m} h_{q+n-m-1} = \sum_{i+j=p+q+n-1} g_i h_j = 0$$

by the preceding equation. Hence $GH = 0$. ■

Theorem 3. Let $C = \langle g(t) \rangle$ be a cyclic code (that is, a nonzero ideal) in the ring B_n and let $1 - x^n = gh$ in $\mathbb{Z}_2[x]$, where $\deg g = m$ and $\deg h = k = n - m$. If G and H are as in (*) and (**), then the word form of the code C is given by

$$C = \{uG \mid u \in B^k\} = \{w \in B^n \mid wH = 0\}.$$

In other words, G is a generator matrix for the word form of the code C , and H is a parity check matrix for the word form of C .

Proof. We already know $C = \{uG \mid u \in B^k\}$. Write $A = \{w \in B^n \mid wH = 0\}$. Then Lemma 2 shows that $C \subseteq A$, so as $|C| = 2^k$ by Theorem 2, it remains to show that $|A| = 2^k$. To this end, let C_0 denote the subspace of B^n spanned by the m columns of H . These columns are independent (in equation (**), $h_k = 1$ because $\deg h = k$), so $\dim C_0 = m$, whence $|C_0| = 2^m$. On the other hand, consider the orthogonal complement C_0^\perp of C_0 , defined by

$$C_0^\perp = \{w \in B^n \mid w \bullet z = 0 \text{ for all } z \in C_0\},$$

where $w \bullet z$ denotes the dot product. If $\{w_1, w_2, \dots, w_m\}$ is a basis of C_0 and $B = [w_1 \ w_2 \ \dots \ w_m]$, then $C_0^\perp = \{w \in B^n \mid wB = 0\} = \{w \in B^n \mid B^T w^T = 0\}$, so a basic theorem of linear algebra shows that

$$\dim C_0^\perp = \dim B^n - \dim C_0 = n - \text{rank } B^T = n - m = k.⁷⁸$$

The proof is completed by the observation that

$$A = \{w \in B^n \mid w \bullet z = 0 \text{ for all columns } z \text{ of } H\} = C_0^\perp. \quad \blacksquare$$

⁷⁸If $\text{null } A = \{X \mid AX = 0\}$, then $\dim(\text{null } A) = n - \text{rank } A$. Note that the usual proof that $\dim C_0^\perp = n - \dim C_0$ breaks down because $w \bullet w = 0$ can happen with $w \neq 0$ ($\text{char } B_n = 2$).

Example 6. We have $1 - x^7 = (1 + x + x^3)(1 + x + x^2 + x^4)$. Hence, take $n = 7$, $m = 3$, and $k = 4$, using $g(t) = 1 + t + t^3 = 1101000$ and $h(t) = 1 + t + t^2 + t^4 = 1110100$. Then

$$G = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

This is the Hamming $(7, 4)$ -code discussed (with a slightly different notation) in Examples 9 and 12, §2.11. \square

Example 7. For any $n \geq 2$, $1 - x^n = (1 + x)(1 + x + x^2 + \dots + x^{n-1})$, and the code $\langle 1 + t \rangle$ consists of the polynomials of even parity. Here, $g(t) = 1 + t = 1100 \dots 0$ and $h(t) = 1 + t + \dots + t^{n-1} = 111 \dots 1$. Hence,

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

In this case, H is obviously a parity check matrix for the code. \square

Error Detection

So far we have paid no attention to the error detecting and correcting capabilities of a cyclic code C . They depend on the minimum distance d of C , that is, by Theorem 4 §2.11, on the minimum weight of a nonzero code word in C . Here, the weight, $\text{wt } c$, of a word c in C is the number of 1's occurring as bits in C . Theorem 4 gives a lower bound on d , which is useful in constructing some important cyclic codes.

Theorem 4 involves the following notion. Let F be any field that contains \mathbb{Z}_2 (for example, any Galois field $GF(2^q)$). Given $n \geq 1$, an element ζ of F is a primitive n th root of unity over \mathbb{Z}_2 if it has order n in the group F^* of nonzero elements of F . Hence, $\zeta^n = 1$ but $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ are all distinct. Note that n must be odd because $n = 2m$ gives $0 = \zeta^n - 1 = (\zeta^m - 1)^2$, so $\zeta^m = 1$. Observe that

$$x^n - 1 = (x - 1)(x - \zeta)(x - \zeta^2) \cdots (x - \zeta^{n-1})$$

because $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ are all roots of $x^n - 1$ in F and they are distinct (ζ is primitive). Hence, every divisor g of $x^n - 1$ (and so the generator of each cyclic code) is a product of terms $(x - \zeta^i)$. In particular, the roots of g in F are all powers of ζ .

Theorem 4. Let $C = \langle g(t) \rangle$ be a cyclic code in B_n . If ζ is a primitive n th root of unity over \mathbb{Z}_2 , assume that t consecutive powers of ζ are roots of g , say

$$g(\zeta^b) = g(\zeta^{b+1}) = \cdots = g(\zeta^{b+t-1}) = 0.$$

Then $d \geq t + 1$, where d is the minimum distance of the code C .

Proof. Let $f(t) = f_0 + f_1t + \cdots + f_{n-1}t^{n-1}$ be an element of C and write the corresponding word as $\bar{f} = f_0f_1 \cdots f_{n-1}$. It suffices to show that $\text{wt } \bar{f} \geq t + 1$. Now $f = qg$ for some q in $\mathbb{Z}_2[x]$ by Theorem 1, so $f(\zeta^{b+i}) = 0$ for $0 \leq i \leq t - 1$. Matrix multiplication gives

$$[f_0 \ f_1 \ \cdots \ f_{n-1}] \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \zeta^b & \zeta^{b+1} & \cdots & \zeta^{b+t-1} \\ \zeta^{2b} & \zeta^{2(b+1)} & \cdots & \zeta^{2(b+t-1)} \\ \vdots & \vdots & & \vdots \\ \zeta^{(n-1)b} & \zeta^{(n-1)(b+1)} & \cdots & \zeta^{(n-1)(b+t-1)} \end{bmatrix} = [0 \ 0 \ \cdots \ 0].$$

Now suppose that $\text{wt } \bar{f} = s \leq t$ so that \bar{f} has exactly s nonzero bits, for instance, $f_{i_1} = f_{i_2} = \cdots = f_{i_s} = 1$ and $f_i = 0$ otherwise. Hence, only rows i_1, i_2, \dots, i_s in the matrix contribute to the product. Consider these rows and the first s columns of the matrix product. The result is

$$[f_{i_1} \ f_{i_2} \ \cdots \ f_{i_s}] \begin{bmatrix} \zeta^{i_1 b} & \zeta^{i_1(b+1)} & \cdots & \zeta^{i_1(b+s-1)} \\ \zeta^{i_2 b} & \zeta^{i_2(b+1)} & \cdots & \zeta^{i_2(b+s-1)} \\ \vdots & \vdots & & \vdots \\ \zeta^{i_s b} & \zeta^{i_s(b+1)} & \cdots & \zeta^{i_s(b+s-1)} \end{bmatrix} = [0 \ 0 \ \cdots \ 0].$$

Hence, this $s \times s$ matrix has zero determinant; that is,

$$0 = \zeta^{i_1 b + i_2 b + \cdots + i_s b} \det \begin{bmatrix} 1 & \zeta^{i_1} & \cdots & (\zeta^{i_1})^{s-1} \\ 1 & \zeta^{i_2} & \cdots & (\zeta^{i_2})^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & \zeta^{i_s} & \cdots & (\zeta^{i_s})^{s-1} \end{bmatrix}.$$

But this is a contradiction because this last determinant is nonzero. (It is a Vandermonde determinant⁷⁹ and $\zeta^{i_1}, \zeta^{i_2}, \dots, \zeta^{i_s}$ are distinct because ζ is primitive.) Hence, $\text{wt } \bar{f} \geq t + 1$, as required. ■

Theorem 4 suggests a way to construct a cyclic code with any predetermined minimum distance d : Just choose a generator polynomial having as roots $d - 1$ consecutive powers of a primitive root of unity. To do so, we recall a notion introduced in Section 6.2. If F is a finite field containing \mathbb{Z}_2 and $v \in F$, the minimal polynomial of v over \mathbb{Z}_2 is the nonzero polynomial m in $\mathbb{Z}_2[x]$ of least degree such that $m(v) = 0$. Then, if $f \in \mathbb{Z}_2[x]$, we have (Theorem 3 §6.2)

$$f(v) = 0 \quad \text{if and only if} \quad m \text{ divides } f \text{ in } \mathbb{Z}_2[x].$$

⁷⁹See, for example, Nicholson, W.K., *Linear Algebra with Applications*, 7th ed., Whitby: PWS-Kent, 2012, Section 3.2.

In particular, minimal polynomials are irreducible, and any irreducible polynomial is the minimal polynomial of each of its roots in \mathbb{Z}_2 .

The code is constructed as follows. Let $2 \leq d \leq n$ and $0 \leq b$ be integers, ζ be a primitive n th root of unity over \mathbb{Z}_2 , and m_i be the minimal polynomial over \mathbb{Z}_2 of ζ^i . If

$$g = \text{lcm}(m_b, m_{b+1}, \dots, m_{b+d-2}),$$

then the cyclic code $\langle g(t) \rangle$ in B_n is called the **binary BCH code**⁸⁰ of length n and **designated distance** d . Note that because the minimal polynomials m_{b+i} are all irreducible, g is the product of the distinct polynomials in the list $m_b, m_{b+1}, \dots, m_{b+d-2}$. Of course, two of these polynomials may be equal.

Theorem 5 collects some basic properties of these BCH codes.

Theorem 5. *Let $C = \langle g(t) \rangle$ be a BCH code as defined above.*

- (1) *The minimum distance of C is at least d .*
- (2) *$f(t) \in C$ if and only if $f(\zeta^i) = 0$ for $i = b, b+1, \dots, b+d-2$.*
- (3) *The following matrix H is a parity check matrix for C :*

$$H = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \zeta^b & \zeta^{b+1} & \cdots & \zeta^{b+d-2} \\ \zeta^{2b} & \zeta^{2(b+1)} & \cdots & \zeta^{2(b+d-2)} \\ \vdots & \vdots & & \vdots \\ \zeta^{(n-1)b} & \zeta^{(n-1)(b+1)} & \cdots & \zeta^{(n-1)(b+d-2)} \end{bmatrix}.$$

Proof. (1) Because m_i divides g for each i , each of $\zeta^b, \zeta^{b+1}, \dots, \zeta^{b+d-2}$ is a root of g . Hence, (1) follows from Theorem 4.

(2) If $f(t)$ is in C , then g divides f , so $g(\zeta^i) = 0$ implies that $f(\zeta^i) = 0$. Conversely, if $f(\zeta^i) = 0$ for each $i = b, b+1, \dots, b+d-2$, then m_i divides f by the definition of m_i , and so g divides f by the definition of g . This shows that $f(t)$ is in C and so proves (2).

(3) Given $f(t) = f_0 + f_1 t + \dots + f_{n-1} t^{n-1}$ in B_n , write the corresponding word as $\bar{f} = f_0 f_1 \dots f_{n-1}$. Then

$$\bar{f}H = [f(\zeta^b) \quad f(\zeta^{b+1}) \quad \cdots \quad f(\zeta^{b+d-2})]$$

so (2) shows that $\bar{f}H = 0$ if and only if $f(t) \in C$, which proves (3). \blacksquare

Example 8. The polynomial $1 + x + x^3$ is irreducible over \mathbb{Z}_2 , so if ζ is one of its roots, we can construct the Galois field $F = GF(8)$ as follows (Theorem 2 §4.3):

$$F = \{a_0 + a_1\zeta + a_2\zeta^2 \mid a_i \in \mathbb{Z}_2, \zeta^3 = 1 + \zeta\}.$$

The powers of ζ in F are $1, \zeta, \zeta^2, \zeta^3 = 1 + \zeta, \zeta^4 = \zeta + \zeta^2, \zeta^5 = 1 + \zeta + \zeta^2, \zeta^6 = 1 + \zeta^2$, and $\zeta^7 = 1$. Thus, ζ is a primitive seventh root of unity over \mathbb{Z}_2 with minimal polynomial $m_1 = 1 + x + x^3$. Moreover, ζ^2 is also a root of m_1 (the third root is ζ^4). Hence, as two consecutive powers of ζ are roots of $m_1(x)$, Theorem 4 guarantees

⁸⁰Discovered by A. Hocquenghem in 1959 and independently by R. C. Bose and D. V. Ray-Chaudhuri in 1960, and hence the name BCH.

that the BCH code $C = \langle 1 + t + t^3 \rangle$ in B_7 has a minimum distance of at least 3 and has a parity check matrix

$$H = \begin{bmatrix} 1 & 1 \\ \zeta & \zeta^2 \\ \vdots & \vdots \\ \zeta^6 & \zeta^{12} \end{bmatrix}.$$

This code is the Hamming (7, 4)-code, and the minimum distance is in fact 3 because $1 + t + t^3$ has weight 3. We described it in Example 6 with a different parity check matrix.

Then minimal polynomial of $\zeta^0 = 1$ is $1 + x$, so the polynomial

$$g = (1 + x)(1 + x + x^3) = 1 + x^2 + x^3 + x^4$$

has three consecutive powers, ζ^0, ζ^1 , and ζ^2 , as roots. Hence, $\langle g(t) \rangle$ has a minimum distance of at least 4 by Theorem 5 (in fact it is 4). Thus, C can detect three errors and correct one error by Theorem 4 §2.11.

Example 9. The polynomial $1 + x + x^4$ is irreducible over \mathbb{Z}_2 (Exercise 11). If ζ is a root, we get the Galois field

$$F = GF(16) = \{a_0 + a_1\zeta + a_2\zeta^2 + a_3\zeta^3 \mid a_i \in \mathbb{Z}, \zeta^4 = 1 + \zeta\}.$$

The powers of ζ are

$$\begin{array}{lll} \zeta^1 = \zeta & \zeta^6 = \zeta^2 + \zeta^3 & \zeta^{11} = \zeta + \zeta^2 + \zeta^3 \\ \zeta^2 = \zeta^2 & \zeta^7 = 1 + \zeta + \zeta^3 & \zeta^{12} = 1 + \zeta + \zeta^2 + \zeta^3 \\ \zeta^3 = \zeta^3 & \zeta^8 = 1 + \zeta^2 & \zeta^{13} = 1 + \zeta^2 + \zeta^3 \\ \zeta^4 = 1 + \zeta & \zeta^9 = \zeta + \zeta^3 & \zeta^{14} = 1 + \zeta^3 \\ \zeta^5 = \zeta + \zeta^2 & \zeta^{10} = 1 + \zeta + \zeta^2 & \zeta^{15} = 1 \end{array}$$

Hence, ζ is a primitive 15th root of unity over \mathbb{Z}_2 . The minimum polynomial of ζ is $m_1 = 1 + x + x^4$, and both ζ^2 and ζ^4 are roots, as is easily verified. The minimal polynomial of ζ^3 is $m_2 = 1 + x + x^2 + x^3 + x^4$ (Exercise 10), so

$$g = m_1 m_2 = 1 + x^4 + x^6 + x^7 + x^8$$

has ζ, ζ^2, ζ^3 , and ζ^4 as roots. Both m_1 and m_2 divide $x^{15} - 1$, so the BCH code

$$C = \langle 1 + t^4 + t^6 + t^7 + t^8 \rangle$$

is a (15, 7)-code, with minimum distance at least 5 by Theorem 5. (The distance is 5 because $1 + t^4 + t^6 + t^7 + t^8$ has weight 5.) Hence, C can detect four errors and correct two errors by Theorem 4 §2.11.

Example 10. Let $F = GF(2^a)$ be the Galois field of order 2^a . Then the multiplicative group F^* of nonzero elements of F is cyclic by Theorem 7 §6.4, say, $F^* = \langle \zeta \rangle$. Writing $n = 2^a - 1$, this means that ζ is a primitive n th root of unity over \mathbb{Z}_2 . Hence, BCH codes with $n = 2^a - 1$ are called *primitive*.

As these examples indicate, Theorem 4 is useful for constructing codes, and these BCH codes are of great practical importance. For example, the European and transatlantic communication system uses a BCH (255, 231)-code that detects

six errors and has a failure rate of 1 in 16 million. As another example, a BCH (128, 112)-code that detects three errors and corrects two errors is used to communicate with the INTELSAT-V satellite.

In addition to Theorem 4, BCH codes are useful because they admit an efficient error-correcting algorithm. A complete discussion of this algorithm is beyond the scope of this book, but we conclude this section with a sketch of the procedure. Given a BCH code C , suppose that a code word c in B^n is transmitted and w is received with errors in bits a_1, a_2, \dots, a_r . Then $w = c + e$, where e is the error word with polynomial form $e = x^{a_1} + x^{a_2} + \dots + x^{a_r}$. The decoder must determine the integers a_j and then decode w by changing bit a_j for each j . Now w is known and so are the quantities $s_i = w(\zeta^i)$, $i = b, b+1, \dots, b+d-2$, where ζ is a primitive n th root of unity over \mathbb{Z}_2 . If H is as in Theorem 5, then $wH = [s_b \ s_{b+1} \ \dots \ s_{b+d-2}]$, so Theorem 5 gives $w \in C$ if and only if $s_i = 0$ for all i . In particular, $cH = 0$, so $e(\zeta^i) = s_i$ for all i because $w = c + e$. Thus,

$$\zeta^{ia_1} + \zeta^{ia_2} + \dots + \zeta^{ia_r} = e(\zeta^i) = s_i, \quad i = b, b+1, \dots, b+d-2.$$

The idea of the decoding algorithm is to determine the quantities ζ^{a_i} from these equations in terms of the (known) s_i . They are determined as the roots of the **error-locator polynomial**:

$$s = (x + \zeta^{a_1})(x + \zeta^{a_2}) \cdots (x + \zeta^{a_r}).$$

Because the roots of s are powers of ζ , we can determine these roots by substituting the powers in s one by one. So the real problem is finding the coefficients of s in terms of the s_i . The main difficulty is that the number of errors r is not known even though algorithms for finding it are known.⁸¹ We content ourselves with an example where $r = 2$.

Example 11. The $(15, 7)$ -code $C = \langle 1 + t^4 + t^6 + t^7 + t^8 \rangle$ in Example 9 can correct two errors. Assume that two errors do occur in bits a and b so that $e = x^a + x^b$. Then the error-locator polynomial is

$$s = (x + \zeta^a)(x + \zeta^b) = x^2 + (\zeta^a + \zeta^b)x + \zeta^{a+b}.$$

Now $\zeta^a + \zeta^b = e(\zeta) = s_1$ is known; to find ζ^{a+b} in terms of the s_i , we compute

$$s_1^3 = (\zeta^a + \zeta^b)^3 = \zeta^{3a} + \zeta^{3b} + \zeta^{a+b}(\zeta^a + \zeta^b) = s_3 + \zeta^{a+b}s_1.$$

Hence, $\zeta^{a+b} = s_1^2 + \frac{s_3}{s_1}$, so the error-locator polynomial is

$$s = x^2 + s_1x + \left(s_1^2 + \frac{s_3}{s_1} \right).$$

To illustrate how this procedure works, suppose that $c = 1 + x^4 + x^6 + x^7 + x^8$ is transmitted and that $w = 1 + x + x^4 + x^6 + x^7$ is received (with errors in bits 1 and 8). Using the formulas in Example 9, we have

$$\begin{aligned} s_1 &= w(\zeta) = 1 + \zeta + \zeta^4 + \zeta^6 + \zeta^7 = 1 + \zeta + \zeta^2, \\ s_3 &= w(\zeta^3) = 1 + \zeta^3 + \zeta^{12} + \zeta^{18} + \zeta^{21} = \zeta. \end{aligned}$$

⁸¹See Williams, F.J. and Sloane, N.J.A., *The Theory of Error-Correcting Codes*, New York: North-Holland, 1977.

Hence, because $s_1^{-1} = \zeta + \zeta^2$ in $GF(16)$, we get $s = x^2 + (1 + \zeta + \zeta^2)x + (\zeta + \zeta^3)$. Then $s(\zeta) = 0 = s(\zeta^8)$, so the roots of $s(x)$ are ζ^1 and ζ^8 , locating errors in bits 1 and 8.

Exercises 6.7

1. (a) Show that $(f_1 + f_2 + \cdots + f_n)^2 = f_1^2 + f_2^2 + \cdots + f_n^2$ for all f_i in $\mathbb{Z}_2[x]$.
 (b) Show that $f(x)^2 = f(x^2)$ for all f in $\mathbb{Z}_2[x]$.
 (c) If $f \in \mathbb{Z}_2[x]$, show that $f' = 0$ if and only if $f = g(x)^2$ for some g in $\mathbb{Z}_2[x]$. Here, f' is the derivative of f .
2. Confirm that $\langle 1 + t \rangle$ is the code of all polynomials in B_n of even parity.
3. Show that the ideals of B_n form a chain if and only if $n = 2^k$ for some $k \geq 1$.
4. Draw the lattice diagram of all codes in B_6 .
5. (a) Find all generator polynomials for the code $C = \langle 1 + t \rangle$ in B_4 .
 (b) Repeat (a) in B_5 . [Hint: Exercise 10 and Theorems 1 and 2.]
6. Let $C = \langle 1 + t \rangle$ and $D = \langle 1 + t + \cdots + t^{n-1} \rangle$ in B_n .
 - (a) If n is odd, show that $C \cap D = \{0\}$ and $B_n \cong C \times D$ as rings.
 - (b) If n is even, show that $D \subseteq C$.
7. How many cyclic codes are there of length

(a) 7	(b) 6	(c) 8
(d) 12	(e) 10	
8. Show that every cyclic code C in B_n has the form $C = \text{ann } f(t)$ for some divisor $f \neq 1$ of $1 - x^n$ in $\mathbb{Z}_2[x]$.
9. Let E be a finite field containing \mathbb{Z}_2 and assume that u is a primitive element for E (that is, $E^* = \langle u \rangle$; see Theorem 7 §6.4). If m_i is the minimal polynomial of u^i , show that m_i divides $x^n - 1$, where $n = |E| - 1$.
10. (a) Show that $1 + x + x^2 + x^3 + x^4$ is irreducible in $\mathbb{Z}_2[x]$.
 (b) Show that $1 + x + x^2 + x^3 + x^4 + x^5 + x^6$ is not irreducible in $\mathbb{Z}_2[x]$. Compare with Example 13 §4.2.
11. Show that $1 + x + x^4$ is irreducible in $\mathbb{Z}_2[x]$.
12. Factor $1 - x^9$ into irreducibles in $\mathbb{Z}_2[x]$.
13. Show that

$$1 - x^{15} = (1 + x)(1 + x + x^2)(1 + x + x^4)(1 + x^3 + x^4)(1 + x + x^2 + x^3 + x^4)$$
 is the factorization of $1 - x^{15}$ into irreducibles in $\mathbb{Z}_2[x]$.
14. Find the generating polynomial for a BCH (31, 16)-code that corrects three errors.
 [Hint: Show that $x^5 + x^2 + 1$ is irreducible in $\mathbb{Z}_2[x]$ and use a root ζ to construct $GF(32)$. Show that $x^5 + x^2 + 1$, $x^5 + x^4 + x^3 + x^2 + 1$, and $x^5 + x^4 + x^2 + x + 1$ are the minimal polynomials of ζ , ζ^3 , and ζ^5 , respectively.]
15. Suppose that a cyclic code C in B_n contains a word of odd parity. Show that $1 + t + t^2 + \cdots + t^{n-1}$ is in C .
16. If n is odd, show that $x^n - 1$ is square free when factored into irreducibles in $\mathbb{Z}_2[x]$.
 [Hint: See the proof of Theorem 3 §6.4.]
17. Assume that $C = \langle g(t) \rangle$ is a cyclic code in B_n .
 - (a) If n is odd, show that $C = \langle e(t) \rangle$, where $e(t)^2 = e(t)$ in B_n . [Hint: By Exercise 16, write $x^n - 1 = gh$, where g and h are relatively prime in $\mathbb{Z}_2[x]$. By Theorem 10 §4.2, write $1 = qg + ph$ in $\mathbb{Z}_2[x]$ and take $e = qg$.]
 (b) Show that $e(t)$ in (a) is uniquely determined by C , called the **idempotent generator** of C .

- (c) Find the idempotent generator for $C = \langle 1 + t + t^3 \rangle$ in B_7 .
 (d) Find the idempotent generator for $C = \langle 1 + t \rangle$ in B_n (n odd).
 (e) If $n = 2^k$, show that B_n contains no idempotents except 0 and 1. [Hint: Exercise 1(b).] Find an idempotent in B_6 other than 0 or 1.
18. Let $C = \langle g(t) \rangle$ in B_n , where g divides $x^n - 1$ in $\mathbb{Z}_2[x]$ and $\deg g = m$. Let u_1, u_2, \dots, u_m be the roots of g in some splitting field $E \supseteq \mathbb{Z}_2$.
- (a) If the roots u_i are distinct, show that the matrix H is a parity check matrix for C (with entries in E).
 (b) If n is odd, show that the roots u_i are necessarily distinct.
19. (Requires elementary linear algebra). Let G be a generator matrix for an (n, k) -code C in B^n . Carry G to row echelon form R by elementary row operations. Show that R has the block form $R = [I_k \ A]$, where A is a $k \times (n - k)$ matrix, and that R is also a generator matrix for C (and so is a *standard* generator matrix for C —see the discussion preceding Theorem 6 §2.11).
20. Let $g = g_0 + g_1x + \dots + g_{n-1}x^{n-1}$ and $h = h_0 + h_1x + \dots + h_{n-1}x^{n-1}$ in $\mathbb{Z}_2[x]$. Show that $g(t)h(t) = 0$ in B^n if and only if $\bar{g} = g_0g_1 \dots g_{n-1}$ is orthogonal to $\bar{h} = h_{n-1}h_{n-2} \dots h_0$ and to every cyclic shift of \bar{h} .

$$H = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ u_1 & u_2 & \cdots & u_m \\ u_1^2 & u_2^2 & \cdots & u_m^2 \\ \vdots & \vdots & & \vdots \\ u_1^{n-1} & u_2^{n-1} & \cdots & u_m^{n-1} \end{bmatrix}$$

Chapter 7

Modules over Principal Ideal Domains

Algebra is generous, she often gives more than is asked of her.

—Jean LeRond d'Alembert

One of the goals of abstract algebra (and of other parts of mathematics for that matter) is to take a class of algebraic structures and show that each object in the class can be systematically constructed from simple and well-understood objects in the class. In this short chapter, we achieve this goal for the class of all finitely generated abelian groups: Each such group is isomorphic to the direct product of a finite number of cyclic groups. In fact, with little extra effort, we actually prove a more general version of this result which has far-reaching implications. This is achieved by introducing the concept of a module which, apart from its intrinsic interest, has become an indispensable tool in several areas of algebra and its applications. In the present case, the abelian groups turn out to be the modules over the ring \mathbb{Z} of integers; our generalization is to look at modules over an arbitrary principal ideal domain. As a by-product, we obtain the classical description of the finitely generated abelian groups as direct products of cyclic groups.

7.1 MODULES

Much of what we say about modules is motivated by abelian groups. It is customary to write abelian groups additively, and we shall do so throughout this chapter. Hence, the unity is called *zero* and denoted 0, the inverse of an element x is called the *negative* of x , denoted $-x$, and an exponent x^n becomes nx for any integer n .

Introduction to Abstract Algebra, Fourth Edition. W. Keith Nicholson.
© 2012 John Wiley & Sons, Inc. Published 2012 by John Wiley & Sons, Inc.

Thus, the equation $x^0 = 1$ in multiplicative notation becomes $0x = 0$ in an additive group. The laws of exponents translate as follows:

Multiplicative Notation	Additive Notation
$g^1 = g$	$1x = x$
$g^{n+m} = g^n g^m$	$(n+m)x = nx + mx$
$(g^m)^n = g^{mn}$	$n(mx) = (nm)x$
$(gh)^n = g^n h^n$ (when $gh = hg$)	$n(x+y) = nx + ny$

In their additive form these rules are exactly the axioms for scalar multiplication in a vector space (see Section 6.1), except that here m and n are restricted to come from the ring \mathbb{Z} . The definition of a module unifies these two cases.

Let M denote an abelian group (written additively). If R is any ring, we say that M is a **left R -module** if, for any $r \in R$ and $x \in M$, an element $rx \in M$ is defined such that the following conditions hold for all $x, y \in M$ and all $r, s \in R$:

- M1 $r(x+y) = rx + ry$
- M2 $(r+s)x = rx + sx$
- M3 $r(sx) = (rs)x$
- M4 $1x = x$

Using M1 and M2, we have $0x = 0$ and $r0 = 0$ for all $x \in M$ and $r \in R$ (Exercise 1).

The multiplication rx where $r \in R$ and $x \in M$ is called the (left) **action** of R on the abelian group M . We will write $_RM$ to indicate that M is a left R -module. Similarly, M is called a right R -module (denoted M_R) if $xr \in M$ for all $x \in M$ and $r \in R$, and the left-right analogues of M1–M4 hold.

Example 1. The \mathbb{Z} -modules are just the (additive) abelian groups.

If $R = F$ is a field, the F -modules are the vector spaces.

Example 2. A ring R becomes an R -module ${}_RR$, where the action is the ring multiplication.

Example 3. Let M_1, M_2, \dots, M_n be R -modules, $n \geq 1$. The set of n -tuples

$$M_1 \oplus M_2 \oplus \cdots \oplus M_n = \{(x_1, x_2, \dots, x_n) \mid x_i \in M_i \text{ for each } i\}$$

is an R -module using *componentwise* addition and R -action:

$$\begin{aligned} (x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) &= (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n) \\ r(x_1, x_2, \dots, x_n) &= (rx_1, rx_2, \dots, rx_n) \end{aligned}$$

for all $r \in R$. This module is called the (**external**) **direct sum** of the M_i . If M is any module, we will write the direct sum of n copies of M as

$$M^n = M \oplus M \oplus \cdots \oplus M.$$

Many of the notions we have been exploring for groups and vector spaces have natural analogues for R -modules. We briefly discuss these one by one.

1. Homomorphisms. Every group homomorphism α preserves exponents in the sense that $\alpha(g^n) = \alpha(g)^n$ for all integers n and all g in the domain of α . In additive

language this is $\alpha(nx) = n\alpha(x)$. Accordingly, if R -modules, we say that a map $\alpha : M \rightarrow N$ is an **R -homomorphism (R -morphism for short)** if:

- H1 $\alpha(x + y) = \alpha(x) + \alpha(y)$, for all $x, y \in M$
- H2 $\alpha(rx) = r\alpha(x)$, for all $x \in M$ and $r \in R$

We also say that α is **R -linear** in this case. Note that every R -morphism is a homomorphism of additive groups by H1. If $R = \mathbb{Z}$ is the ring of integers, then H2 follows from H1, so the \mathbb{Z} -morphisms are precisely the group homomorphisms. For a field $R = F$, the F -morphisms are just the linear transformations.

If R -module, the **trivial morphism** $\alpha : R\text{-}M \rightarrow R\text{-}N$ (where $\alpha(x) = 0$ for all $x \in M$) is always R -linear, as is the **identity morphism** $1_M : M \rightarrow M$ given by $1_M(x) = x$ for each $x \in M$. Given modules $R\text{-}M$, $R\text{-}N$, and $R\text{-}K$, and R -morphisms $M \xrightarrow{\beta} N \xrightarrow{\alpha} K$, the composite $\alpha\beta : M \rightarrow K$ is also R -linear as is easily verified.

If an R -morphism is one-to-one and onto, it is called an **R -isomorphism**. Two modules $R\text{-}M$ and $R\text{-}N$ are called **isomorphic** (written $M \cong N$) if there exists an R -isomorphism $\sigma : M \rightarrow N$. In this case $\sigma^{-1} : N \rightarrow M$ is also R -linear (verify) and so is also an R -isomorphism. The results in Theorem 3 §2.5 and its Corollary remain true for R -morphisms in general.

2. Submodules. If $R\text{-}M$ is an R -module, a subset $N \subseteq M$ will be called an **R -submodule** if the following conditions are satisfied:

- S1 N is a subgroup of the (additive, abelian) group M
- S2 rx is in N for all $r \in R$ and $x \in N$

Thus, the \mathbb{Z} -submodules of a group are just the subgroups and, if F is a field, the F -submodules of a vector space are the subspaces. If M is a module, both M and $\{0\}$ are submodules of M ; we call $\{0\}$ the **zero submodule** and write $\{0\} = 0$. Note that a submodule N of M is an R -module in its own right, that is M1–M4 hold for N (verify). Moreover, submodules of N are already submodules of M .

Example 4. The submodules of $R\text{-}R$ are called the **left ideals** of the ring R ; they are the additive subgroups L of R for which $rx \in L$ for all $r \in R$ and $x \in L$. Similarly the **right ideals** of R are the submodules of R_R . Thus, Lemma 2 §3.3 asserts that the ideals of R are precisely the additive subgroups of R that are simultaneously right and left ideals. Note that in a commutative ring the sets of ideals, left ideals, and right ideals coincide.

Example 5. Let $R\text{-}M$ be an R -module. If A is a left ideal of R and X is a nonempty subset of M , define AX to be the set of all finite sums of elements ax , $a \in A$, $x \in X$:

$$AX = \{a_1x_1 + a_2x_2 + \cdots + a_nx_n \mid n \geq 1, a_i \in A \text{ and } x_i \in X\}.$$

Then AX is a submodule of $R\text{-}M$ as is easily verified. In particular, if $x \in M$ then $Rx = \{rx \mid r \in R\}$ is a submodule of M called the **principal submodule** generated by x . Note that if $R = \mathbb{Z}$ then $\mathbb{Z}x$ is the subgroup generated by x .

3. Kernels and Images. If $\alpha : R\text{-}M \rightarrow R\text{-}N$ is an R -linear mapping then α is a group homomorphism and so has a **kernel** and an **image** as defined in Section 2.5:

- (1) $\ker \alpha = \{x \in M \mid \alpha(x) = 0\}$.
- (2) $\operatorname{im} \alpha = \{\alpha(x) \mid x \in M\} = \alpha(M)$.

Moreover, these are submodules of M and N , respectively. In fact, S1 follows in both cases because α is a group homomorphism (Theorem 1 §2.10). To prove S2 for $\ker \alpha$, let $x \in \ker \alpha$ and observe that, if $r \in R$ then $rx \in \ker \alpha$ because $\alpha(rx) = r\alpha(x) = r0 = 0$. Similarly, if $\alpha(x) \in \alpha(M)$ then $r\alpha(x) = \alpha(rx) \in \alpha(M)$ for all $r \in R$, so S2 holds for $\alpha(M)$ too.

If $\alpha : M \rightarrow N$ is R -linear, it is clear that α is onto if and only if $\text{im } \alpha = N$. Moreover, since α is also \mathbb{Z} -linear, Theorem 3 §2.10 shows that α is one-to-one if and only if $\ker \alpha = 0$. We will use these facts frequently below.

4. Factor Modules. Let N be a submodule of $_RM$. Then N is a subgroup of the (abelian) additive group M , so N is normal and the factor group M/N exists. Writing $x + N$ for the additive coset generated by x , the factor group takes the form $M/N = \{x + N \mid x \in M\}$. This becomes an R -module via the action

$$r(x + N) = rx + N, \quad \text{for all } r \in R \text{ and } x \in M.$$

To see that this is well defined, let $x + N = y + N$ in M/N ; we must show that $rx + N = ry + N$ for any $r \in R$. Since $x + N = y + N$, we have $x - y \in N$ by Theorem 1 §2.6. Hence $rx - ry = r(x - y) \in N$ because N is a submodule, and so $rx + N = ry + N$ (again by Theorem 1 §2.6). This is what we wanted.

With this, it is a routine matter to verify M1-M4 for the factor group M/N . Thus M/N is an R -module, called the **factor module**. Moreover, the coset map

$$\varphi : M \rightarrow M/N \quad \text{given by } \varphi(x) = x + N \text{ for all } x \in M$$

is an onto R -morphism with $\ker \varphi = N$. As an illustration, the map $k \mapsto \bar{k}$ from $\mathbb{Z} \rightarrow \mathbb{Z}_n$ is a \mathbb{Z} -morphism so $a\bar{k} = \bar{ak}$ for all $a, k \in \mathbb{Z}$.

If $\alpha : _RM \rightarrow _RN$ is R -linear then Theorem 4 §2.10 shows that there exists a \mathbb{Z} -isomorphism $\sigma : M/\ker \alpha \rightarrow \alpha(M)$ given by $\sigma(x + \ker \alpha) = \alpha(x)$ for all x in M . This map σ actually satisfies H2 (verify), so σ is R -linear and hence is an R -isomorphism. This proves

Theorem 1. Module Isomorphism Theorem. If $\alpha : _RM \rightarrow _RN$ is R -linear then $M/\ker \alpha \cong \alpha(M)$.

As one illustration, let $_RM$ be a module, let $x \in M$, and consider the map $\theta : R \rightarrow Rx$ defined by $\theta(r) = rx$ for all $r \in R$. This is R -linear and onto, and $\ker \theta = \{r \in R \mid rx = 0\}$. This is a left ideal of R called the **annihilator** of x , and denoted $\text{ann } x = \{r \in R \mid rx = 0\}$. Hence, the isomorphism theorem gives

Corollary 1. If $M = _RM$ is a module and $x \in M$, then $Rx \cong R/\text{ann } x$.

An element x in a module $_RM$ is called **torsion-free** if $\text{ann } x = 0$, that is if $rx = 0$, $r \in R$, implies that $r = 0$. Hence $Rx \cong _RR$ in this case. Clearly 0 is never torsion free; while in a vector space *every* nonzero element is torsion free. In an abelian group the torsion free elements are precisely the elements of infinite order.

Accordingly, we say that x is a **torsion** element if $\text{ann } x \neq 0$. Hence, in an abelian group the torsion elements are just the elements of finite order. We say that $_RM$ is a **torsion module** (a **torsion-free module**) if every nonzero element is torsion (torsion free). We will have more to say about these modules below.

If K and N are submodules of a module M , then $K \cap N$ is also a submodule as is $K + N = \{k + n \mid k \in K, n \in N\}$. The proof of the following consequences of Theorem 1 is Exercise 7.

Corollary 2. Let K and N be submodules of a module M . Prove

- (1) **Second Isomorphism Theorem.** $(K + N)/K \cong N/(K \cap N)$.
- (2) **Third Isomorphism Theorem.** If $K \subseteq N$ then $M/N \cong (M/K)/(N/K)$.

One very useful method of describing a module M is to show that it is an external direct sum of well-understood modules. There is a way to do this “internally”, working only with submodules of M , that will be used extensively below.

Direct Sums

If H_1, \dots, H_m are submodules of a module $_RM$ their **sum** $H_1 + \dots + H_m$ is defined to be the set of all finite sums of elements of the H_i , more formally

$$H_1 + \dots + H_m = \{x_1 + \dots + x_m \mid m \geq 1, x_i \in H_i \text{ for each } i\}.$$

This is a submodule of M . In fact it is the *smallest* submodule containing every H_i in the sense that any submodule containing all the H_i must contain $H_1 + \dots + H_m$.

For example, if a and b are integers, the set $X = \{ra + sb \mid r, s \in \mathbb{Z}\}$ of all linear combinations of a and b played a prominent role in the study of \mathbb{Z} in Section 1.2. Now we see that $X = Ra + Rb$ is the sum of the ideals of \mathbb{Z} generated by a and b .

Before proceeding, we note an important property of sums of submodules.

Theorem 2. Modular Law. Let K , N , and H be submodules of a module M . If $K \subseteq N$ then $N \cap (K + H) = K + (N \cap H)$.

Proof. Clearly $K + (N \cap H) \subseteq N \cap (K + H)$. Conversely, let $n \in N \cap (K + H)$, say $n = k + h$ with the obvious notation. Then $h = n - k \in N \cap H$ because $K \subseteq N$, so $n = k + h \in K + (N \cap H)$. This proves that $N \cap (K + H) \subseteq K + (N \cap H)$. ■

Let H_1, \dots, H_m be submodules of some module. Clearly every element of the sum $H_1 + \dots + H_m$ is a sum of elements of the H_i , but in general this can happen in more than one way. We say that $H_1 + \dots + H_m$ is a **direct sum** if this representation is *unique* for each $x \in H_1 + \dots + H_m$; that is, given two representations,

$$x_1 + \dots + x_m = x = y_1 + \dots + y_m$$

where $x_i \in H_i$ and $y_i \in H_i$ for each i , then necessarily $x_i = y_i$ for each i .

Theorem 3. The following conditions are equivalent for a set H_1, \dots, H_m of submodules of a module M :

- (1) The sum $H_1 + \dots + H_m$ is direct.
- (2) If $x_1 + \dots + x_m = 0$ where $x_i \in H_i$ for each i , then each $x_i = 0$.
- (3) $(\sum_{i \neq k} H_i) \cap H_k = 0$ for each k .
- (4) $(H_1 + \dots + H_{k-1}) \cap H_k = 0$ for each $k \geq 2$.

In this case $H_1 + \dots + H_m \cong H_1 \oplus \dots \oplus H_m$ —the **external direct sum**.

Proof. (1) \Rightarrow (2). We have $x_1 + \dots + x_m = 0 = 0 + \dots + 0$, so each $x_i = 0$ by (1).

(2) \Rightarrow (3). Let $x \in (\Sigma_{i \neq k} H_i) \cap H_k$, so that $x = \Sigma_{i \neq k} x_i$ with $x_i \in H_i$ for each $i \neq k$. Then $x_1 + \cdots + x_{k-1} + (-x) + x_{k+1} + \cdots + x_m = 0 = 0 + \cdots + 0 + \cdots + 0$. Hence $-x = 0$ by (2), so $x = 0$ proving (3).

(3) \Rightarrow (4). This is clear since $H_1 + \cdots + H_{k-1} \subseteq \Sigma_{i \neq k} H_i$ for each $k \geq 2$.

(4) \Rightarrow (1). Let $x_1 + \cdots + x_m = y_1 + \cdots + y_m$ be two representations of an element in $H_1 + \cdots + H_m$, where $x_i, y_i \in H_i$ for each i . Then $(x_1 - y_1) + \cdots + (x_m - y_m) = 0$ and $x_i - y_i \in H_i$ for each i . Then $x_m - y_m \in (\Sigma_{i < m} H_i) \cap H_m = 0$ by (4), so $x_m = y_m$. Continuing in this way, (4) implies that $x_i = y_i$ for each $i < m$.

Finally, define $\sigma : H_1 \oplus \cdots \oplus H_m \rightarrow H_1 + \cdots + H_m$ by

$$\sigma(x_1, \dots, x_m) = x_1 + \cdots + x_m, \quad \text{where } x_i \in H_i \text{ for each } i.$$

Then σ is R -linear and onto, and it is one-to-one by (2). So σ is an isomorphism. \blacksquare

Corollary 1. If $K, N \subseteq M$ then $K + N$ is direct if and only if $K \cap N = 0$.

Hence if $M = K \oplus N$ then $M/K \cong N$ and $M/N \cong K$ (Exercise 9).

Example 6. Let $m = pq$ in \mathbb{Z} , where $\gcd(p, q) = 1$. Show that $\mathbb{Z}_m = \mathbb{Z}\bar{p} \oplus \mathbb{Z}\bar{q}$.

Solution. If $k \in \mathbb{Z}$, let $\bar{k} \in \mathbb{Z}_m$ denote the residue class. Let $1 = xp + yq$, $x, y \in \mathbb{Z}$ (as $\gcd(p, q) = 1$). Then $\bar{k} = kxp + kyq = kx\bar{p} + ky\bar{q}$ for all $k \in \mathbb{Z}$, so $\mathbb{Z}_m = \mathbb{Z}\bar{p} + \mathbb{Z}\bar{q}$.

To see that $\mathbb{Z}\bar{p} \cap \mathbb{Z}\bar{q} = \{\bar{0}\}$, let $\bar{k} \in \mathbb{Z}\bar{p} \cap \mathbb{Z}\bar{q}$, say $\bar{k} = a\bar{p} = b\bar{q}$ where $a, b \in \mathbb{Z}$. Then $a\bar{p} = b\bar{q}$ so $m|(ap - bq)$. Since $q|m$ it follows that $q|ap$. But then $q|a$ (again because $\gcd(p, q) = 1$), say $a = zq$ with $z \in \mathbb{Z}$. Hence, $\bar{k} = a\bar{p} = z\bar{q}\bar{p} = z\bar{m} = z\bar{0} = \bar{0}$. This shows that $\mathbb{Z}\bar{p} \cap \mathbb{Z}\bar{q} = \{\bar{0}\}$, as required by Corollary 1 of Theorem 3. \square

Example 7. If $e^2 = e \in R$ is an idempotent, show that $R = Re \oplus R(1 - e)$.

Solution. We have $R = Re + R(1 - e)$ because $1 \in Re + R(1 - e)$ and $Re + R(1 - e)$ is a left ideal. Now suppose that $a \in Re \cap R(1 - e)$, say $a = re = s(1 - e)$, $r, s \in R$. Then $ae = re^2 = [s(1 - e)]e = s(e - e^2) = 0$ because $e = e^2$. It follows that $a = 0$, which proves that $Re \cap R(1 - e) = 0$. Hence, $R = Re \oplus R(1 - e)$ by Corollary 1 of Theorem 3, as required. \square

Note that the converse of Example 7 is also true—see Exercise 15.

Let H_1, H_2, \dots, H_m be submodules of a module M such that $H_1 + H_2 + \cdots + H_m$ is a direct sum. In view of the isomorphism in the last sentence of Theorem 3, it is customary to abuse the language somewhat and write

$$H_1 + H_2 + \cdots + H_m = H_1 \oplus H_2 \oplus \cdots \oplus H_m.$$

We refer to this as the **internal direct sum** of the H_i . The fact that we use the same notation for internal and external direct sums causes little confusion: It is nearly always clear from the context which one is intended.

Corollary 2. Let $M = H_1 \oplus H_2 \oplus \cdots \oplus H_m$ be a direct sum of modules. If $M = K_1 + K_2 + \cdots + K_m$ and $K_i \subseteq H_i$ are submodules, then $K_i = H_i$ for each i .

Proof. Let $x_i \in H_i$ for each i and write $x = x_1 + x_2 + \cdots + x_m$. By hypothesis, let $x = k_1 + k_2 + \cdots + k_m$, where $k_i \in K_i \subseteq H_i$ for each i . Because $H_1 \oplus \cdots \oplus H_m$ is direct, $k_i = x_i$ for each i by Theorem 3, so $x_i \in K_i$. Hence, $H_i \subseteq K_i$ for each i . \blacksquare

If $M = H_1 \oplus H_2 \oplus \cdots \oplus H_m$ is an internal direct sum of modules, it is impossible to overemphasize the importance of the uniqueness of the representation of each

element $x \in M$ in the form $x = h_1 + h_2 + \cdots + h_m$ with $h_i \in H_i$ for each i . This is evident in the proof of Corollary 2. As another illustration, if $K_i \subseteq H_i$ is a submodule for each i , consider the map

$$\varphi : M \rightarrow \frac{H_1}{K_1} \oplus \frac{H_2}{K_2} \oplus \cdots \oplus \frac{H_m}{K_m}$$

given by $\varphi(h_1 + h_2 + \cdots + h_m) = (h_1 + K_1, h_2 + K_2, \dots, h_m + K_m)$. Then φ is well defined because of the uniqueness. Hence φ is an onto module homomorphism, as is easily verified, and $\ker \varphi = K_1 \oplus K_2 \oplus \cdots \oplus K_m$ (this sum is direct because sums from the K_i are sums from the H_i). The isomorphism theorem now gives

Corollary 3. Let $M = H_1 \oplus H_2 \oplus \cdots \oplus H_m$, and let $K_i \subseteq H_i$ be a submodule for each i . Then $K = K_1 \oplus K_2 \oplus \cdots \oplus K_m$ is a direct sum and $\frac{M}{K} \cong \frac{H_1}{K_1} \oplus \frac{H_2}{K_2} \oplus \cdots \oplus \frac{H_m}{K_m}$.

The following result will be used several times below. The proof is Exercise 12.

Corollary 4. Suppose $M = H_1 \oplus H_2 \oplus \cdots \oplus H_m$ and $N = K_1 \oplus K_2 \oplus \cdots \oplus K_m$. If $H_i \cong K_i$ for each i , then $M \cong N$.

Free Modules

Up to this point the concepts we have discussed about R -modules are analogues of those for abelian groups (the \mathbb{Z} -modules). Free modules reflect the vector spaces.

Given elements x_1, \dots, x_k in a module $_RM$, a sum $r_1x_1 + \cdots + r_kx_k$, $r_i \in R$, is called a **linear combination** of the x_i , with **coefficients** r_i . The sum of the submodules Rx_i is a submodule

$$Rx_1 + \cdots + Rx_k = \{r_1x_1 + \cdots + r_kx_k \mid r_i \in R\},$$

consisting of *all* such linear combinations. This is the smallest submodule of M that contains every x_i , and so is called the submodule **generated** by the x_i . If $M = Rx_1 + \cdots + Rx_k$ we say that $\{x_1, \dots, x_k\}$ is a **generating set** for M . If M has a generating set of n elements we say that M is n -generated; M is said to be **finitely generated** if it is n -generated for some n . Note that the finitely generated vector spaces are just the finite dimensional ones.

As for vector spaces, a set $\{x_1, \dots, x_k\}$ in $_RM$ is called **independent** if

$$r_1x_1 + \cdots + r_kx_k = 0, r_i \in R, \text{ implies that } r_1 = \cdots = r_k = 0,$$

that is, if the only linear combination that vanishes is the **trivial** one with every coefficient zero. Pursuing the vector space analogy, a subset $\{x_1, \dots, x_k\}$ is called a **basis** of M if it is independent and generates M . The following characterization of these bases will be needed.

Theorem 4. The following conditions are equivalent for $\{x_1, \dots, x_k\} \subseteq _RM$:

- (1) $\{x_1, \dots, x_k\}$ is a basis of M .
- (2) $M = Rx_1 \oplus \cdots \oplus Rx_k$ and each x_i is torsion-free.

Proof. (1) \Rightarrow (2). Assume (1). Then $M = Rx_1 + \cdots + Rx_k$ because the x_i generate M . If $r_1x_1 + \cdots + r_kx_k = 0$, $r_i \in R$, then each $r_i = 0$ by independence. So certainly $r_ix_i = 0$ for each i , which shows that $Rx_1 + \cdots + Rx_k$ is a direct sum. But if $rx_i = 0$

for some i then $0x_1 + \cdots + rx_i + \cdots + 0x_k = 0$, so $r = 0$, again by independence. This shows that each x_i is torsion free, and so proves (2).

(2) \Rightarrow (1). The x_i generate M because $M = Rx_1 \oplus \cdots \oplus Rx_k$. Suppose that $r_1x_1 + \cdots + r_kx_k = 0$, $r_i \in R$. Then each $r_ix_i = 0$ because $Rx_1 \oplus \cdots \oplus Rx_k$ is direct, so each $r_i = 0$ because the x_i are torsion free. Hence, the x_i are independent. ■

A module that has a finite basis is called a **free module**.⁸² Every finite dimensional vector space is free (it has a finite basis by Theorem 6 §6.1); however, in general finitely generated modules need not contain a basis (for example, the \mathbb{Z} -module \mathbb{Z}_n has no basis because it has no torsion-free elements). In fact, if R is a domain *every* free module is torsion free (Exercise 25). On the other hand, for a PID we will show that the converse holds for finitely generated free modules (Corollary to Theorem 1 §7.2). The “finitely generated” condition is essential: The set \mathbb{Q} of rational numbers is a torsion-free \mathbb{Z} -module that is not free (Exercise 25).

Example 8. R^n is free with basis $\{(1, 0, \dots, 0), (0, 1, \dots, 0), (0, 0, \dots, 1)\}$, called the **standard basis** of R^n —see Example 11 §6.1.

As for vector spaces, R -morphisms $RW \rightarrow RM$ are easy to define if RW is free.

Theorem 5. Let RW be free with a basis $\{w_1, \dots, w_n\}$. Given any module RM and arbitrary elements x_1, \dots, x_n in M , define

$$\theta : W \rightarrow M \quad \text{by} \quad \theta(\sum r_i w_i) = \sum r_i x_i,$$

where the r_i are in R . Then θ is an R -linear map such that $\theta(w_i) = x_i$ for each i .

Proof. The map θ is well defined because each element $w \in W$ has the form $w = \sum r_i w_i$ (the w_i generate W), and this representation is unique (the w_i are independent). The rest is a routine verification. ■

Every image of a finitely generated module is again finitely generated (verify). Corollary 1 below is a useful converse of this: If M in Theorem 5 is generated by $\{x_1, \dots, x_n\}$, then the map θ is onto and we obtain

Corollary 1. Every n -generated module is an image of a free module with a basis of n elements.

If the module M in Theorem 5 is itself free with basis $\{x_1, \dots, x_n\}$, then θ is onto because the x_i generate M , and θ is one-to-one because the x_i are independent. Hence θ is an isomorphism and this, together with Example 8, gives

Corollary 2. Every free module with a basis of n elements is isomorphic to R^n .

Corollary 3. If P and Q are free with bases of m and n elements, respectively, then $P \oplus Q$ is free with a basis of $m + n$ elements.

Proof. Regard $P \oplus Q$ as an internal direct sum. If $\{x_1, \dots, x_m\}$ and $\{y_1, \dots, y_n\}$ are bases of P and Q , respectively, then $\{x_1, \dots, x_m, y_1, \dots, y_n\}$ is a basis of $P \oplus Q$ as the reader can verify. ■

⁸²There is a general definition of a free module. Let X be a possibly infinite set in a module M . Then X is called a basis of M if it is independent (each finite subset of X is independent) and generates M (every element of M is a linear combination of (finitely many) elements of X). Then a module is called free if it has a basis. Our concern is about finite bases, but much of what we do generalizes.

If RW is free with basis $\{w_1, \dots, w_n\}$, another important special case of Theorem 5 arises as follows: For each $k \in \{1, 2, \dots, n\}$ define

$$\pi_k : W \rightarrow R \quad \text{by} \quad \pi_k(\sum r_i w_i) = r_k.$$

These are onto, R -linear maps for each k , called the **projections** associated with the basis $\{w_1, \dots, w_n\}$. The following useful property is easily verified

$$x = \sum_k \pi_k(x) x_k, \quad \text{for all } x \in W.$$

In contrast to Corollary 1, a free module W can only be the image of a module M if M contains a direct summand isomorphic to W . A module P is called **projective** if it satisfies the following property:

$$\text{If } \alpha : M \rightarrow P \text{ is } R\text{-linear and onto then } M = \ker \alpha \oplus P_1$$

for some submodule P_1 of M . Note that $P_1 \cong P$; indeed $P_1 \cong M/\ker \alpha \cong \alpha(M) = P$. Projective modules are important in ring theory and homological algebra.

Theorem 6. *Every free module with a finite basis is projective.*

Proof. Let $\alpha : M \rightarrow W$ be onto and R -linear. If $\{w_1, \dots, w_n\}$ is a basis of W choose $x_i \in M$ such that $\alpha(x_i) = w_i$ (α is onto). By Theorem 5, there exists $\theta : W \rightarrow M$ such that $\theta(w_i) = x_i$ for each i . It follows that $\alpha[\theta(w_i)] = \alpha(x_i) = w_i$ for each i , and hence that $\alpha(\theta(w)) = w$ for every $w \in W$ (since the w_i generate W). In other words, $\alpha\theta = 1_W$. With this we can show

$$M = \ker \alpha \oplus \theta(W).$$

We have $\alpha[x - \theta\alpha(x)] = \alpha(x) - \alpha\theta\alpha(x) = 0$ for all $x \in M$, so $M = \ker \alpha + \theta(W)$ because $x = [x - \theta\alpha(x)] + \theta\alpha(x)$. But if $y \in \ker \alpha \cap \theta(W)$ write $y = \theta(w)$, where $w \in W$. Then $w = 1_W(w) = \alpha\theta(w) = \alpha(y) = 0$, so $y = \theta(w) = \theta(0) = 0$. This completes the proof. ■

Up to this point the properties of modules we have derived are valid for all rings R , even noncommutative ones, but sometimes restrictions on R must be made. For example, the invariance theorem (Theorem 5 §6.1) shows that any two bases of a finite dimensional vector space have the same number of elements (the dimension), but this theorem fails for free modules in general. We say that the ring R has **invariant basis number (IBN)** if, whenever a free R -module has a finite basis, then any two bases have the same number of elements. In this language, the above discussion shows that all fields have IBN. The next theorem shows that it is enough that R has an image R/A that is a field. In fact it is enough that R/A has IBN for some ideal A of R .

Suppose A is an ideal of R , and $W = RW$ is a module. Then AW is a submodule of W (Example 5) and we claim that W/AW is a module over the ring R/A with the action $(r+A)(w+AW) = rw+AW$ for all $r \in R$ and $w \in W$. To see that this is well-defined, let $r+A = s+A$ and $w+AW = v+AW$; we must show that $rw+AW = sv+AW$. But $rw-sv = r(w-v) + (r-s)v \in AW$ because $v-w \in AW$ and $r-s \in A$. The module axioms are routine verifications. Now we can prove

Theorem 7. *Let R be a ring that has an ideal A such that R/A has IBN. Then R has IBN.*

Proof. If $\{w_i \mid 1 \leq i \leq k\}$ is a basis of RW , it is enough to show that the set $\{w_i + AW \mid 1 \leq i \leq k\}$ is a basis of $R/A(W/AW)$ —then k is uniquely determined by hypothesis. Given $w + AW$ in W/AW , write $w = \sum_i r_i w_i$, $r_i \in R$. Then

$$w + AW = (\sum_i r_i w_i) + AW = \sum_i r_i (w_i + AW) = \sum_i (r_i + A)(w_i + AW),$$

which shows that $\{w_i + AW \mid 1 \leq i \leq k\}$ generates W/AW . To see that it is independent, observe first that $AW = \bigoplus_i Aw_i$ (Exercise 21). With this, suppose that $\sum_i (r_i + A)(w_i + AW) = 0$, $r_i \in R$. Then $\sum_i r_i w_i + AW = 0$, so $\sum_i r_i w_i \in AW$, say $\sum_i r_i w_i = \sum_i a_i w_i$, $a_i \in A$. As the w_i are independent this implies that $r_i = a_i \in A$ for each i , so $r_i + A = 0$ for each i . This shows that $\{w_i + AW \mid 1 \leq i \leq k\}$ is independent in $R/A(W/AW)$, and so is a basis. This is what we wanted. ■

If R has IBN, the number of elements in any basis of a free module W is called the **rank** of W , and denoted $\text{rank } W$. Thus, rank and dimension are the same for vector spaces. The ring \mathbb{Z} of integers has IBN because $\mathbb{Z}/\langle p \rangle = \mathbb{Z}_p$ is a field for every prime p . More generally, every PID R has IBN. Indeed: If R is a field there is nothing to do; otherwise R contains a prime element p (it is a UFD by Theorem 2 §5.2) and $R/\langle p \rangle$ is a field by Theorem 3 §5.2. Hence

Corollary. Every PID has IBN.

Remark. In fact *every* commutative ring has IBN, but the proof involves a set-theoretic result called **Zorn's lemma** (see Appendix C). This guarantees the presence of a maximal ideal A in *any* ring R (even noncommutative), and R/A is a field if R is commutative.

Exercises 7.1

Throughout these exercises R always denotes a ring.

1. If M is an R -module, show that:
 - (a) $0x = 0$ for all $x \in M$.
 - (b) $r0 = 0$ for all $r \in R$.
 - (c) $(-1)x = -x$ for all $x \in M$.
2. Let $K \subseteq M$ be modules. Show that
 - (a) If M is n -generated, then every image of M is m -generated for $m \leq n$.
 - (b) If $M = K \oplus N$ is finitely generated, so are K and N .
 - (c) If both K and M/K are finitely generated, so is M .
3. If $e^2 = e \in R$, show that Re is an ideal of R if and only if $eR(1 - e) = 0$.
4. Let $R = F \times F \times \dots$, (F a field), a ring with componentwise operations, and let $A = \{(a_1, a_2, \dots) \mid a_i = 0 \text{ for all but finitely many } i\}$. Show that A is an ideal of R that is not finitely generated.
5. Let A and B be left ideals of R , and let K and N be submodules of RM .
 - (a) Show that AK is a submodule of M .
 - (b) Show that $A(K + N) = AK + AN$.
 - (c) Show that $(A + B)K = AK + BK$.
6. Let A be an ideal of R , and let RM be a module. If both $_RA$ and M are finitely generated, show that AM is finitely generated. [Hint: Exercise 5.]
7. Let K and N be submodules of a module M .
 - (a) Show that $(K + N)/K \cong N/(K \cap N)$.
 - (b) If $K \subseteq N$ show that N/K is a submodule of M/K and $(M/K)/(N/K) \cong M/K$.

8. Let R be an integral domain. Given $_RM$ let $T(M) = \{t \in M \mid t \text{ is torsion}\}$.
- Show that $T(M)$ is a submodule of M —called the **torsion submodule**.
 - Show that $T[M/T(M)] = 0$. We say that $M/T(M)$ is **torsion-free**.
9. If $M = P \oplus Q$ are modules, show that $M/P \cong Q$ and $M/Q \cong P$. [Hint: Theorem 3(2).]
10. Let $K \subseteq N$ be submodules of a module M . If $K \cap X = N \cap X$ and $K + X = N + X$ for some submodule X . Show that $K = N$. [Hint: $N = N \cap (N + X)$.]
11. Let $M = \mathbb{Z} \oplus \mathbb{Z}$, and $K = \{(k, k) \mid k \in \mathbb{Z}\}$. Determine if $M = K \oplus X$ in case
- $X = \{(k, 0) \mid k \in \mathbb{Z}\}$
 - $X = \{(0, k) \mid k \in \mathbb{Z}\}$
 - $X = \{(2k, 3k) \mid k \in \mathbb{Z}\}$
 - What if $X = \{(k, -k) \mid k \in \mathbb{Z}\}$?
12. If $M_i \cong N_i$, $1 \leq i \leq k$, show that $M_1 \oplus M_2 \oplus \cdots \oplus M_k \cong N_1 \oplus N_2 \oplus \cdots \oplus N_k$.
13. Let $M = M_1 \oplus M_2 \oplus \cdots \oplus M_r$ be an internal direct sum of modules. If we have $M_1 = K_1 \oplus \cdots \oplus K_s$, show that $M = K_1 \oplus \cdots \oplus K_s \oplus M_2 \oplus \cdots \oplus M_r$.
14. Let $M = M_1 \oplus M_2 \oplus \cdots \oplus M_r$ be an internal direct sum of modules. If $K_i \subseteq M_i$ for each i , show that $K_1 + K_2 + \cdots + K_r$ is a direct sum.
15. If $R = A \oplus B$ where A and B are left ideals, show that there exists $e^2 = e \in R$ such that $A = Re$ and $B = R(1-e)$. [Hint: Let $1 = e + f$, $e \in A$, $f \in B$. If $a \in A$ consider $a - ae = af \in A \cap B$.]
16. Given $_RM$, an R -linear map $\pi : M \rightarrow M$ is called a **projection** if $\pi^2 = \pi$.
- If π is a projection, show that $M = \pi(M) \oplus \ker \pi$.
 - If $M = N \oplus K$, find a projection π such that $N = \pi(M)$ and $K = \ker \pi$.
17. Let $M \xrightarrow{\beta} N \xrightarrow{\alpha} M$ be R -linear. If $\alpha\beta = 1_M$, show that $N = \beta(M) \oplus \ker \alpha$.
18. If $M = \mathbb{Z}_2 \oplus \mathbb{Z}_4$, write $K = \mathbb{Z}_2 \oplus 0$, $N = 0 \oplus \mathbb{Z}_4$, and $X = 0 \oplus \{0, 2\}$. Show that
- $K \cong X$ but $M/K \not\cong M/X$.
 - $M/(K \oplus X) \cong M/N$ but $K \oplus X \not\cong N$.
19. If G is an abelian group and $|G| = mn$, where $\gcd(m, n) = 1$, show that $G = G_m \oplus G_n$, where $G_k = \{g \in G \mid kg = 0\}$.
20. If R is a domain, show that every finitely generated free R -module is torsion free.
21. Let A be an ideal of a ring R , let $_RW$ be a module, and consider the R/A -module W/AW with action $(r+A)(w+AW) = rw+AW$ as in the discussion preceding Theorem 7. If $\alpha : W \rightarrow {}_RV$ is R -linear, define $\bar{\alpha} : W/AW \rightarrow V/AV$ by $\bar{\alpha}(w+AW) = \alpha(w)+AV$ for all $w+AW \in W/AW$.
- Show that $\bar{\alpha}$ is well defined.
 - Show that $\bar{\alpha}$ is (R/A) -linear.
22. A module $_RM$ is called **simple** if 0 and M are the only submodules. If R is a ring, show that ${}_RR$ is simple if and only if R is a division ring.
23. If $_RM$ and ${}_RN$ are simple (preceding exercise), prove **Schur's Lemma**: If $\alpha : M \rightarrow N$ is R -linear, then either $\alpha = 0$ or α is an isomorphism.
24. Show that the following conditions on a finitely generated module P are equivalent:
- P is projective.
 - P is isomorphic to a direct summand of a free module.
 - If α, β are R -linear and α is onto in the diagram,
then γ exists such that $\alpha\gamma = \beta$.
 - If $\alpha : M \rightarrow P$ is onto and R -linear,
there exists $\gamma : P \rightarrow M$ such that $\alpha\gamma = 1_P$.

[Hint: For (2) \Rightarrow (3) assume that $F = P \oplus Q$ is free for some module Q . Define $\pi : F \rightarrow P$ by $\pi(p+q) = p$ for all $p \in P$ and $q \in Q$. If $\{x_1, \dots, x_n\}$ is a basis of F choose $m_i \in M$ such that $\alpha(m_i) = \beta\pi(x_i)$ for each i . By Theorem 5, there exists an R -homomorphism $\theta : F \rightarrow M$ such that $\theta(x_i) = m_i$ for each i .]

$$\begin{array}{ccc} & & P \\ & \swarrow \gamma & \downarrow \beta \\ M & \xrightarrow{\alpha} & N \end{array}$$

25. Show that \mathbb{Q} is a torsion-free \mathbb{Z} -module that is not free.

[Hint: Call an additive group Q divisible if, for all $0 < n \in \mathbb{Z}$ and all $q \in Q$, the equation $nx = q$ has a solution $x \in \mathbb{Q}$. Show that \mathbb{Q} is divisible, direct summands of divisible groups are divisible, and \mathbb{Z} is not divisible.]

7.2 MODULES OVER A PID

Unless otherwise noted, throughout this section R will denote a principal ideal domain (PID); that is R is an integral domain and every ideal A of R has the form $A = Ra$ for some $a \in R$.⁸³ These rings are discussed in detail in Section 5.2. The main example is $R = \mathbb{Z}$, in which case the modules are the abelian groups (written additively). We will state most of the theorems in this section in the case that R is a general PID, but the reader should keep the abelian group case in mind for motivation.

The goal of this section is to completely describe the finitely generated modules over a PID, and the following theorem is fundamental.

Theorem 1. *If R is a PID, every finitely generated module $_RM$ has a decomposition as a direct sum of principal submodules Rx_i , $x_i \in M$:*

$$M = Rx_1 \oplus Rx_2 \oplus \cdots \oplus Rx_n.$$

Due to its difficulty, the proof of Theorem 1 (and of a uniqueness property) is left to the end of this section; we focus instead on how Theorem 1 is used to give explicit information about finitely generated modules. In particular, we obtain a complete description of all finitely generated abelian groups.

It is routine to verify that if M is a free module over a domain with a finite basis then M is finitely generated and torsion free. If the ring is a PID the converse holds. Recall that if M is a free module with a finite basis over a PID, the number of elements in the basis is uniquely determined (Corollary to Theorem 5 §7.1) called the rank of M .

Theorem 2. *If R is a PID then a module $_RM$ is free of finite rank if and only if it is finitely generated and torsion-free.*

Proof. If M is free of rank n , let $\{w_1, \dots, w_n\}$ be any basis. Then $M = \sum R w_i$ so M is certainly finitely generated. If $am = 0$ with $0 \neq m \in M$, write $m = \sum r_i w_i$, $r_i \in R$. Then $0 = aw = \sum (ar_i)w_i$ so each $ar_i = 0$ because the w_i are independent. But some $r_i \neq 0$ (because $w \neq 0$), it follows that $a = 0$ (because R is a domain). Hence M is torsion free.

Conversely, if M is finitely generated and torsion-free, then Theorem 1 shows that $M = Rx_1 \oplus Rx_2 \oplus \cdots \oplus Rx_n$. We may assume that $x_i \neq 0$ for all i . Hence, each x_i is torsion free because M is torsion free, so $\{x_1, x_2, \dots, x_n\}$ is a basis of M by Theorem 4 §7.1. ■

The principal submodules Rx arising in Theorem 1 can be easily described in terms of the ring R . The map $r \mapsto rx$ from $R \rightarrow Rx$ is R -linear and onto with

⁸³We use the suggestive notation Ra rather than $\langle a \rangle$ for the ideal generated by $a \in R$.

kernel $\text{ann } x = \{r \in R \mid rx = 0\}$ —called the *annihilator* of x . Hence, $Rx \cong R/\text{ann } x$ by the isomorphism theorem (Theorem 1 §7.1). There are two cases:

If x is torsion free, $\text{ann } x = 0$ so $Rx \cong R$ is free of rank 1.

If x is torsion, $\text{ann } x = Rd$ for some $0 \neq d \in R$ (R is a PID) so $Rx \cong R/Rd$.

In the case that x is torsion, it is instructive to look at the case when $R = \mathbb{Z}$. Then the torsion elements x are just those of finite order. In this case, the order $o(x) = n$ if and only if $\text{ann } x = \mathbb{Z}n$, and this gives a way to extend the notion of order to torsion elements of any module over an arbitrary PID.

Torsion Submodule

Let R be a PID and let $_RM$ be a module. If $x \in M$ is a torsion element then $\text{ann } x$ is a nonzero ideal of R so, as R is a PID, $\text{ann } x = Rd$ for some $0 \neq d \in R$. We define the **order** $o(x)$ of a torsion element x as follows:

$$o(x) = d, \quad \text{where } \text{ann } x = Rd.^{84}$$

Of course $d = o(x)$ is only unique up to multiplication by a unit of R by Theorem 1 §5.1. However, we do have the following properties familiar from the group case: If $o(x) = d \neq 0$ then

- (1) $dx = 0$.
- (2) If $r \in R$, then $rx = 0$ if and only if $d|r$.

The routine verifications are left to the reader.

If $R = \mathbb{Z}$, let x be a torsion element in some group. If $\text{ann } x = \mathbb{Z}d$, $d \neq 0$, then also $\text{ann } x = \mathbb{Z}(-d)$. However, these are the only generators of $\text{ann } x$ (since 1 and -1 are the only units in \mathbb{Z}). Hence, the convention in group theory is to make the order of x unique by choosing the *positive* generator for $\text{ann } x$.

If R is a PID and M is an R -module, the set of all torsion elements of M is a submodule of M , called the **torsion submodule** of M , denoted

$$T(M) = \{x \in M \mid \text{ann } x \neq 0\} = \{x \in M \mid o(x) \neq 0\}.$$

Hence M is torsion free if and only if $T(M) = 0$, and M is torsion if and only if $T(M) = M$. Thus finite abelian groups are torsion as \mathbb{Z} -modules.

Example 1. Let \mathbb{C}^* be the (multiplicative) group of all nonzero complex numbers. Then $T(\mathbb{C}^*)$ consists of all roots of unity. Note that $T(\mathbb{C}^*)$ is an example of a torsion group that is not finite.

Theorem 3. Let M be a finitely generated module over a PID. Then

- (1) $T(M)$ is a torsion submodule of M and $M/T(M)$ is torsion free.
- (2) $M = T(M) \oplus W$, where W is free of finite rank.

Proof. For convenience write $T(M) = T$.

(1) To see that M/T is torsion-free, let $x + T \neq 0$ in M/T ; we must show that $\text{ann}(x + T) = 0$. If not, let $r(x + T) = 0$, where $r \neq 0$. Then $rx \in T$, say $s(rx) = 0$ for some $s \neq 0$ in R . But $sr \neq 0$ because R is a domain, so this implies that $x \in T$, contradicting the assumption that $x + T \neq 0$.

⁸⁴**Warning.** This notion of order does *not* coincide with the group-theoretic notion for torsion-free elements. In a group, x has *infinite* order if and only if $\text{ann}(x) = 0$.

(2) The coset map $\varphi : M \rightarrow M/T$ is onto and $\ker \varphi = T$. Moreover, M/T is finitely generated (since M is) and torsion free (by (1)), and so is free of finite rank by Theorem 2. But then $M = \ker \varphi \oplus W$ for some submodule $W \subseteq M$ by Theorem 6 §7.1. Since $W \cong M/\ker \varphi \cong M/T$, this proves (2). ■

If R is a PID, the finitely generated free modules are well understood (they are all isomorphic to R^n where n is the rank). Thus, Theorem 3 shows that the task of describing all finitely generated modules is reduced to looking at the torsion modules. We now turn to this task.

Primary Decomposition

Let R be a PID, and recall that $p \in R$ is called a **prime** if, whenever $p|ab$ in R then either $p|a$ or $p|b$. If p is a prime and $_RM$ is an R -module define

$$M(p) = \{x \in M \mid p^k x = 0 \text{ for some integer } k \geq 0\}.$$

One verifies that this is a submodule of M for any prime p , called the **p -primary component** of M . Note that in a PID the only divisors of p^k (up to unit multiples) are powers of p . Hence,

$$M(p) = \{x \in M \mid o(x) = p^k \text{ for some integer } k \geq 0\}.$$

Example 2. Write $\mathbb{Z}_{24} = \{0, 1, 2, \dots, 23\}$. Then one verifies that

$$\mathbb{Z}_{24}(2) = \mathbb{Z}_{24}3 = \{0, 3, 6, 9, 12, 15, 18, 21\} \quad \text{and} \quad \mathbb{Z}_{24}(3) = \mathbb{Z}_{24}8 = \{0, 8, 16\}.$$

We have $\mathbb{Z}_{24}(p) = \{0\}$ for all primes p other than 2 or 3.

If M is torsion free then $M(p) = 0$ for all primes p . However, if M is torsion and finitely generated, this is far from the case. Indeed, let $M = Rx_1 + \dots + Rx_k$, where R is a PID. If $o(x_i) = d_i \neq 0$ for each i then $d = d_1 d_2 \dots d_k \neq 0$ and $dM = 0$, where we write $rM = \{rx \mid x \in M\}$. Hence, $\text{ann}(M) \neq 0$, where

$$\text{ann}(M) = \{r \in R \mid rM = 0\}$$

is called the **annihilator** of M . This is an ideal of R , and M is torsion if and only if $\text{ann}(M) \neq 0$.

Theorem 4. Primary Decomposition Theorem. Let R be a PID, let $_RM \neq 0$ be a finitely generated, torsion module, and suppose $dM = 0$ where $0 \neq d \in R$. Since R is a UFD, let $d = p_1^{k_1} p_2^{k_2} \dots p_m^{k_m}$, where the p_i are nonassociated primes in R and each $k_i > 0$. Then

- (1) $M = M(p_1) \oplus M(p_2) \oplus \dots \oplus M(p_m)$.
- (2) If $\text{ann}(M) = Rd$ then $M(p_i) \neq 0$ for each i .

Proof. (1) We begin by showing that $M = M(p_1) + M(p_2) + \dots + M(p_m)$. If $m = 1$ then $M(p_1) = M$ because $dM = 0$. So assume that $m \geq 2$.

In this case write $d_i = d/p_i^{k_i}$ for each i , and let $c = \gcd(d_1, d_2, \dots, d_m)$. We claim that no prime p divides c (if $p|c$ then $p|d_i$ for each i , and this is impossible because $m \geq 2$). Since R is a UFD, this means that c is a unit in R . But Theorem 1 §5.2 shows that $c \in Rd_1 + Rd_2 + \dots + Rd_m$, and it follows that $1 = r_1 d_1 + r_2 d_2 + \dots + r_m d_m$

for some $r_i \in R$. Hence, if $x \in M$ we have

$$x = r_1 d_1 x + r_2 d_2 x + \cdots + r_m d_m x, \quad r_i \in R.$$

Moreover, $p_i^{k_i} (r_i d_i x) = r_i (p_i^{k_i} d_i) x = r_i dx = 0$, so $r_i d_i x \in M(p_i)$ for each i . Hence,

$$M = M(p_1) + M(p_2) + \cdots + M(p_m).$$

To see that this sum is direct, let $x \in [M(p_1) + M(p_2) + \cdots + M(p_{k-1})] \cap M(p_k)$; we must show that $x = 0$ by Theorem 3 §7.1. Suppose $x_1 + x_2 + \cdots + x_{k-1} = x$ with $x_i \in M(p_i)$ for each $i = 1, 2, \dots, k-1$ and $x \in M(p_k)$, say $p_i^{t_i} x_i = 0$ for each i , and $p_k^{t_k} x = 0$. Write $q = p_1^{t_1} \cdots p_{k-1}^{t_{k-1}}$. Then $qx_i = 0$ for each $i < k$. But $\gcd(q, p_k^{t_k}) = 1$ because the p_i are nonassociated, say $1 = rq + sp_k^{t_k}$ where $r, s \in R$. Thus

$$x = 1x = rqx + sp_k^{t_k} x = rqx + 0 = rq(x_1 + x_2 + \cdots + x_{k-1}) = 0,$$

as required. Hence, $M = M(p_1) \oplus M(p_2) \oplus \cdots \oplus M(p_m)$, which proves (1).

(2) Assume that $\text{ann}(M) = Rd$. Each $M(p_i)$ is an image of M by (1), and so is finitely generated. Hence, there exists $s_i \geq 1$ such that $p_i^{s_i} M(p_i) = 0$. Now suppose that $M(p_1) = 0$. This is impossible if $m = 1$ (we are assuming that $M \neq 0$). If $m \geq 2$, and we write $b = p_2^{s_2} \cdots p_m^{s_m}$, it follows that $b \in \text{ann}(M) = Rd$. But then $d|b$, a contradiction because p_1 does not divide b . So $M(p_1) \neq 0$. A similar argument shows that $M(p_i) \neq 0$ for each i . ■

Corollary 1. If R is a PID, let M and N be finitely generated, torsion R -modules. Then $M \cong N$ if and only if $M(p) \cong N(p)$ for all primes $p \in R$.

Proof. If $M(p) \cong N(p)$ for each prime p , then $M \cong N$ by Theorem 4. Conversely, if $\alpha : M \rightarrow N$ is R -linear then $\alpha[M(p)] \subseteq N(p)$, with equality if α is onto (verify). So, if α is an isomorphism then $\alpha : M(p) \rightarrow N(p)$ is an isomorphism. ■

Example 3. Find the primary decomposition of the abelian group $G = \mathbb{Z}_{60}$, and find a generator for each primary component.

Solution. Write $\bar{k} = k$ in G . We have $\text{ann}(G) = \mathbb{Z}60$, so $d = 60 = 2^2 \cdot 3 \cdot 5$ and the primary decomposition is $\mathbb{Z}_{60} = G(2) \oplus G(3) \oplus G(5)$. Moreover, we claim that $G(2) = \mathbb{Z}15$, $G(3) = \mathbb{Z}20$, and $G(5) = \mathbb{Z}12$.

We show that $G(3) = \mathbb{Z}20$, and leave the (similar) verifications of the others to the reader. We have $G(3) = \{a \mid 3^k a = 0 \text{ for some } k \geq 0\}$. Hence $20 \in G(3)$, so $\mathbb{Z}20 \subseteq G(3)$. Conversely, if $a \in G(3)$ then $3^k a = 0$ for some k , whence $60|3^k a$. Hence $4|3^k a$ so, since $\gcd(4, 3^k) = 1$ we have $4|a$. Similarly $5|a$ so, since $\gcd(4, 5) = 1$, it follows that $4 \cdot 5 = 20|a$. Hence $a = 20k$ for some k , so $a \in \mathbb{Z}20$. This proves that $G(3) \subseteq \mathbb{Z}20$. □

A finite abelian group G is a finitely generated, torsion \mathbb{Z} -module. If $p \in \mathbb{Z}$ is a prime, G is called a **p -group** if the order of every element of G is a power of p . Thus, the primary component $G(p) = \{x \in G \mid o(x) = p^k \text{ for some } k \geq 0\}$ is a p -group for each prime p dividing $|G|$. Moreover, $G(p)$ contains every p -subgroup of G and so is the unique largest p -subgroup of G . Thus in Example 3, $\text{ann}(\mathbb{Z}_{60}) = 60$ so $\mathbb{Z}_{60} = \mathbb{Z}_{60}(2) \oplus \mathbb{Z}_{60}(3) \oplus \mathbb{Z}_{60}(5)$. Observe that 60 has prime factorization $60 = 2^2 \cdot 3 \cdot 5$, and $|\mathbb{Z}_{60}(2)| = 2^2$, $|\mathbb{Z}_{60}(3)| = 3$, and $|\mathbb{Z}_{60}(5)| = 5$ by Example 2. This holds for any finite abelian group.

Corollary 2. Primary Decomposition Theorem for Finite Abelian Groups. Let G be a finite abelian group of order $|G| = p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$, where the p_i are distinct primes. Then

$$G = G(p_1) \oplus G(p_2) \oplus \cdots \oplus G(p_m).$$

Moreover, $|G(p_i)| = p_i^{n_i}$ for each i .

Surprisingly, the proof of the last statement in Corollary 2 requires the following important fact.

Lemma 1. Let G be a finite abelian group, and let $p \in \mathbb{Z}$ be a prime.

- (1) If p divides $|G|$, then G has an element of order p .⁸⁵
- (2) G is a p -group if and only if $|G| = p^n$ for some $n \geq 0$.

Proof. (1) Use induction on $|G|$. It is clear if $|G| = 1, 2$, or 3 . If $|G| > 3$, choose some $h \in G$, $h \neq 0$, and write $o(h) = n$. If $p|n$, then $o(\frac{n}{p}h) = p$ and we are done. So assume that $\gcd(p, n) = 1$. If we write $H = \langle h \rangle$, then $|G| = |H| |G/H| = n|G/H|$, so p divides $|G/H|$. By induction, let $g + H$ be a coset in G/H of order p . We claim that $o(pg) = p$. We have $p(g + H) = 0$, so $pg \in H$. Because $|H| = n$, this gives $0 = n(pg) = p(pg)$ by Lagrange's theorem. As p is a prime, it remains to show that $pg \neq 0$. But $pg = 0$ implies that $n(g + H) = 0$ in G/H and so, because $g + H$ has order p , it yields $p|n$, contrary to assumption.

(2) If G is a p -group then (1) shows that p is the only prime divisor of $|G|$, and so $|G| = p^n$ for some $n \geq 0$. The converse is by Lagrange's theorem. ■

Proof of Corollary 2. Write $d = p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$. Then $dM = 0$ by Lagrange's theorem so Theorem 4 shows that $G = G(p_1) \oplus G(p_2) \oplus \cdots \oplus G(p_m)$. Clearly $G(p_i)$ is a p_i -group for each i , so let $|G(p_i)| = p_i^{k_i}$ for some k_i by Lemma 1. But then

$$|G| = |G(p_1)| \cdot |G(p_2)| \cdots |G(p_k)| = p_1^{k_1} p_2^{k_2} \cdots p_m^{k_m}.$$

Since $|G| = d = p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$, it follows that $k_i = n_i$ for each i by the uniqueness of the prime factorization of integers. ■

p -Modules

The primary decomposition theorem shows that to describe the torsion modules M over a PID R it is enough to describe $M(p)$ for each prime $p \in R$. To this end, an R -module M is called a **p -module** if, for each $x \in M$, $p^k x = 0$ for some $k \geq 0$, equivalently if $o(x) = p^n$ for some integer $n \geq 0$. Hence, if $p \in \mathbb{Z}$ is a prime then the p -modules are just the p -groups defined above. Note that $M(p)$ is a p -module for any M , and that images and submodules of p -modules are again p -modules. The following theorem gives a concise description of all p -modules over any PID.

We will need one fact: If p is a prime in a PID R , then the factor ring R/Rp is a field. This is part of Theorem 3 §5.2, and actually characterizes the primes in R .

Theorem 5. Let R be a PID, let $p \in R$ be a prime, and let M be a finitely generated, nonzero p -module over R . Then there is a decomposition

$$M = Rx_1 \oplus Rx_2 \oplus \cdots \oplus Rx_t, \tag{*}$$

⁸⁵This actually holds for *any* finite group, abelian or not, and is called Cauchy's theorem. We prove it in Section 8.2.

where $o(x_i) = p^{m_i}$ with $m_1 \geq m_2 \geq \dots \geq m_t \geq 1$. Furthermore, the integers t, m_1, m_2, \dots, m_t are uniquely determined by M .

More generally, if $K \subseteq M$ is any submodule and $K = Ry_1 \oplus Ry_2 \oplus \dots \oplus Ry_u$ where $o(y_i) = p^{k_i}$ with $k_1 \geq k_2 \geq \dots \geq k_u \geq 1$, then $u \leq t$ and $k_i \leq m_i$ for each $i = 1, 2, \dots, u$.

Proof. The decomposition in (*) exists by Theorem 1, each $o(x_i)$ is a power of p because M is a p -module, and we can ensure that $m_1 \geq m_2 \geq \dots \geq m_t$ by relabeling the x_i . The uniqueness of t and the m_i follows from the last sentence of the theorem with $K = M$.

So let $K \subseteq M$ and, since K is also a p -module, use (*) to write K as in the theorem. We begin by showing that $u \leq t$. Define a submodule $L_p(M)$ of M by

$$L_p(M) = \{x \in M \mid px = 0\}.$$

This is a vector space over the field R/Rp via the action $(r + Rp)x = rx$ for all $r \in R$. Moreover, a routine computation shows that $L_p(Rx_i) = R(p^{m_i-1}x_i)$ for each i , so $L_p(Rx_i)$ has dimension 1 over R/Rp . But

$$L_p(M) = L_p(Rx_1) \oplus L_p(Rx_2) \oplus \dots \oplus L_p(Rx_t),$$

so $L_p(M)$ has R/Rp -dimension t . Hence, $u \leq t$ because $L_p(K) \subseteq L_p(M)$.

The rest is proved by induction on $n \geq 0$, where $p^nM = 0$ (such an n exists because M is a finitely generated p -module). Now $m_i \leq n$ as $p^n \in \text{ann } Rx_i = Rp^{m_i}$, and similarly $k_j \leq n$. If $n = 0$ then $M = 0$ and there is nothing to prove. If $n = 1$ then $k_i = m_i = 1$ for each $i = 1, 2, \dots, u$. If $n \geq 2$ consider $pM = \{px \mid x \in M\}$. This is an R -submodule of M and one verifies that

$$pM = Rp_{\lambda}x_1 \oplus \dots \oplus Rp_{\lambda}x_{\lambda}, \quad \text{where } m_{\lambda} > 1 \text{ and } m_{\lambda+1} = \dots = m_t = 1,$$

$$pK = Rp_{\mu}y_1 \oplus \dots \oplus Rp_{\mu}y_{\mu}, \quad \text{where } k_{\mu} > 1 \text{ and } k_{\mu+1} = \dots = k_u = 1.$$

Now observe that $o(px_i) = p^{m_i-1}$ for $1 \leq i \leq \lambda$ and $o(py_i) = p^{k_i-1}$ for $1 \leq i \leq \mu$. Since $p^{n-1}(pM) = 0$, induction gives $\mu \leq \lambda$ and $k_i \leq m_i$ for $1 \leq i \leq \mu$. But if $\mu < i \leq u$ then $k_i = 1 \leq m_i$, which completes the proof. ■

Let M be a finitely generated, nonzero p -module and, as in Theorem 5, let

$$M = Rx_1 \oplus Rx_2 \oplus \dots \oplus Rx_t,$$

where $o(x_i) = p^{m_i}$ for each i and $m_1 \geq m_2 \geq \dots \geq m_t \geq 1$. Then

The t -tuple (m_1, m_2, \dots, m_t) is called the **type** of the module M ;

The elements $p^{m_1}, p^{m_2}, \dots, p^{m_t}$ are called the **elementary divisors** of M .

The integers m_i and the elementary divisors p^{m_i} are uniquely determined by M .

Given a sequence of integers $m_1 \geq m_2 \geq \dots \geq m_t$ the module

$$R/Rp^{m_1} \oplus R/Rp^{m_2} \oplus \dots \oplus R/Rp^{m_t}$$

is of type (m_1, m_2, \dots, m_t) . Hence, Theorem 5 gives

Corollary 1. Up to isomorphism, there is exactly one finitely generated p -module of each type.

If K and M are two finitely generated p -modules of types (k_1, k_2, \dots, k_u) and (m_1, m_2, \dots, m_t) , respectively, we say that K has **smaller type** than M if $u \leq t$ and $k_i \leq m_i$ for each $i = 1, 2, \dots, u$.

Corollary 2. If M is a finitely generated, nonzero p -module, then

- (1) Every nonzero submodule of M has smaller type.
- (2) M has a submodule of each smaller type.

Proof. (1) This is by Theorem 5.

(2) Let M have type (m_1, m_2, \dots, m_t) , say $M = Rx_1 \oplus \dots \oplus Rx_t$, $o(x_i) = p^{m_i}$ for each i . Suppose that (k_1, k_2, \dots, k_u) is a smaller type. Then the submodule $K = R(p_1^{m_1-k_1}x_1) \oplus \dots \oplus R(p_u^{m_u-k_u}x_t)$ is of type (k_1, k_2, \dots, k_u) . \blacksquare

Theorem 9 §2.4 shows that, if G is a cyclic group of order n , then G has exactly one subgroup of order d for each divisor d of n . Corollary 2 shows that, for finite abelian p -groups, the subgroups, although not absolutely unique as in the cyclic case, are uniquely determined up to type. For example, if $G = \mathbb{Z}_p \oplus \mathbb{Z}_p$, then $K_1 = \mathbb{Z}_p \oplus \{0\}$, $K_2 = \{0\} \oplus \mathbb{Z}_p$, and $K_3 = \{(a, a) \mid a \in \mathbb{Z}_p\}$ all have type (1).

Example 4. If G is a p -group of type $(2, 1, 1)$, the possible types of nonzero subgroups of G are $(2, 1, 1)$, $(1, 1, 1)$, $(2, 1)$, $(1, 1)$, (2) , and (1) . \square

Corollary 3. Let G be a p -group with $|G| = p^n$. If G has type (m_1, m_2, \dots, m_t) , then $n = m_1 + m_2 + \dots + m_t$.⁸⁶

Proof. Let $G = \mathbb{Z}x_1 \oplus \mathbb{Z}x_2 \oplus \dots \oplus \mathbb{Z}x_t$, where $o(x_i) = p^{m_i}$ for each i . Then $|G| = |\mathbb{Z}x_1| |\mathbb{Z}x_2| \dots |\mathbb{Z}x_t| = p^{m_1} p^{m_2} \dots p^{m_t}$, and the result follows. \blacksquare

If G is a cyclic group then $G \cong \mathbb{Z}_m$ or $G \cong \mathbb{Z}$ according as $|G| = m$ or $|G| = \infty$. It is customary to use \mathbb{Z}_m and \mathbb{Z} as representatives of the cyclic groups.

Example 5. Classify the abelian groups of order p^5 , where p is a prime.

Solution. The various types are listed together with a representative group.

Type	Group
(5)	\mathbb{Z}_{p^5}
(4, 1)	$\mathbb{Z}_{p^4} \oplus \mathbb{Z}_p$
(3, 2)	$\mathbb{Z}_{p^3} \oplus \mathbb{Z}_{p^2}$
(3, 1, 1)	$\mathbb{Z}_{p^3} \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$
(2, 2, 1)	$\mathbb{Z}_{p^2} \oplus \mathbb{Z}_{p^2} \oplus \mathbb{Z}_p$
(2, 1, 1, 1)	$\mathbb{Z}_{p^2} \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$
(1, 1, 1, 1, 1)	$\mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$

If $n \geq 1$, all abelian groups of order p^n can be described in the same way. \square

Theorems 4 and 5 provide a way to describe all finite abelian groups. This is demonstrated in the following two examples.

⁸⁶Decompositions $n = m_1 + m_2 + \dots + m_t$ with $m_1 \geq m_2 \geq \dots \geq m_t \geq 1$ are called *partitions* of the integer n and are important in number theory.

Example 6. Describe the abelian groups of order p^2q^3 , where p and q are distinct primes.

Solution. If $|G| = p^2q^3$, then $G = G(p) \oplus G(q)$, where $|G(p)| = p^2$ and $|G(q)| = q^3$ by the Corollary 2 of Theorem 4. Thus, the possible types for $G(p)$ are (2) and $(1, 1)$, whereas those for $G(q)$ are (3) , $(2, 1)$, and $(1, 1, 1)$. Hence, up to isomorphism, there are six abelian groups G of order p^2q^3 :

$$\begin{array}{ll} \mathbb{Z}_{p^2} \oplus \mathbb{Z}_{q^3} & \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_{q^3} \\ \mathbb{Z}_{p^2} \oplus \mathbb{Z}_{q^2} \oplus \mathbb{Z}_q & \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_{q^2} \oplus \mathbb{Z}_q \\ \mathbb{Z}_{p^2} \oplus \mathbb{Z}_q \oplus \mathbb{Z}_q \oplus \mathbb{Z}_q & \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_q \oplus \mathbb{Z}_q \oplus \mathbb{Z}_q \end{array}$$

□

Example 7. How many distinct abelian groups are there of order 1,333,584?

Solution. Because $1,333,584 = 2^4 \cdot 3^5 \cdot 7^3$, the primary components have orders 2^4 , 3^5 , and 7^3 . The various types are

- 2-component $(4), (3, 1), (2, 2), (2, 1, 1), (1, 1, 1, 1)$
- 3-component $(5), (4, 1), (3, 2), (3, 1, 1), (2, 2, 1), (2, 1, 1, 1), (1, 1, 1, 1, 1)$
- 7-component $(3), (2, 1), (1, 1, 1)$

Thus, there are 5, 7, and 3 choices, respectively, for the primary components and hence $5 \cdot 7 \cdot 3 = 105$ choices in all. Theorem 5 and Corollary 1 of Theorem 4 guarantee that no two are isomorphic. □

We have illustrated the general results about modules over a PID using abelian groups (\mathbb{Z} -modules). However, there is another very important example. If F is a field, much of linear algebra is concerned with determining the nature of a linear transformation $\alpha : V \rightarrow V$ where V is a vector space over F . Theorem 1 gives a satisfactory answer when V is finite dimensional. There are two key observations: First, if t is an indeterminant over F then the polynomial ring $F[t]$ is a PID by Theorem 1 §4.3, and each nonzero ideal has a unique monic generator. The second observation is that, given a linear transformation $\alpha : V \rightarrow V$, the vector space V becomes a module over the polynomial ring $F[t]$ via the action

$$pv = p(\alpha)(v), \quad \text{for all } p \in F[t] \text{ and all } v \in V,$$

where we remember that $p(\alpha) : V \rightarrow V$ is a linear transformation for each α . If FV is finite dimensional, one shows that $F[t]V$ is torsion and finitely generated. Hence, the decomposition of $F[t]V$ in Theorems 4 and 5 provide an elegant way to prove many of the basic theorems of linear algebra. Moreover, if A is an $n \times n$ matrix in $M_n(F)$ then this yields the canonical forms for the matrix A by taking $V = F^n$ and considering the linear transformation $\alpha : V \rightarrow V$ given by $\alpha(v) = Av$ for all $v \in V$. However, a discussion of the details is beyond the scope of this book.

The Fundamental Theorem

If R is a PID, we are going to prove Theorem 1 that every finitely generated torsion module over R is a direct sum of principal submodules in a unique way. The whole thing depends of a result about free modules called the submodule theorem. This in turn requires two preliminary results, each of interest in itself.

If \mathcal{S} is a nonempty set of submodules of a module M , a submodule $K \in \mathcal{S}$ is called **maximal** in \mathcal{S} if, whenever $K \subseteq N$ with $N \in \mathcal{S}$, then necessarily $K = N$. For example, the maximal ideals of a ring R are the maximal members of the set $\mathcal{S} = \{A \mid A \text{ is an ideal and } A \neq R\}$.

Lemma 2. *If R is a PID then every nonempty set \mathcal{S} of ideals of R has a maximal member.*

Proof. Assume that \mathcal{S} has no maximal member, and choose $A_1 \in \mathcal{S}$. Then A_1 is not maximal so let $A_1 \subset A_2$, where $A_2 \in \mathcal{S}$. But A_2 is not maximal either, so let $A_1 \subset A_2 \subset A_3$ where $A_3 \in \mathcal{S}$. This process continues to create a strictly increasing sequence $A_1 \subset A_2 \subset A_3 \subset \dots \subset A_k \subset A_{k+1} \subset \dots$ of ideals of R .⁸⁷ Define $A = A_1 \cup A_2 \cup A_3 \cup \dots$. Then A is an ideal so, since R is a PID, let $A = Rd$ where $d \in R$. Since $d \in A$, we have $d \in A_k$ for some $k \geq 1$, and it follows that $A \subseteq A_k \subseteq A_{k+1} \subseteq \dots A$. But this implies that $A_k = A_{k+1}$, a contradiction. ■

Lemma 3. *Let RW be a free module of rank n over a PID R . If $K \neq 0$ is a submodule of W then K is also free and $\text{rank}(K) \leq n$.*

Proof. Since $\text{rank } W = n$ there is an isomorphism $\sigma : W \rightarrow R^n$, so $K \cong \sigma(K) \subseteq R^n$. Hence, we may assume that $K \subseteq R^n$, and we proceed by induction on n . If $n = 1$ then $K \subseteq R$ and the result follows from the fact that each nonzero ideal of R has the form Rd , where d is torsion free. If $n \geq 2$, define $\varepsilon : K \rightarrow R$ by $\varepsilon(r_1, \dots, r_n) = r_1$ whenever $(r_1, \dots, r_n) \in K$. If $\varepsilon(K) = 0$ then $K \subseteq R^{n-1}$ and we are done by induction. If $\varepsilon(K) \neq 0$ then, since it is an ideal of R , write $\varepsilon(K) = Ra$ where $0 \neq a \in R$. Hence $\varepsilon : K \rightarrow Ra$ is onto so, since Ra is free (a is torsion free), Theorem 7 §7.1 shows that $K = \ker \varepsilon \oplus K_1$ where $K_1 \cong Ra$. Now $\ker \varepsilon \subseteq R^{n-1}$, so $\ker \varepsilon$ is free of rank at most $n - 1$ by induction. Since $K_1 \cong Ra$ is free of rank 1, it follows that $K = \ker \varepsilon \oplus K_1$ is free of rank at most n . ■

Important as Lemma 3 is, we need more detailed information about how the free submodule K is positioned in W . This is provided by the following fundamental theorem.

Theorem 6. Submodule Theorem. *Let R be a PID, and let RW be a free module of finite rank n . If $K \neq 0$ is a submodule of W , there exists a basis $\{y_1, y_2, \dots, y_n\}$ of W , an integer $m \leq n$, and nonzero elements d_1, d_2, \dots, d_m of R such that*

- (1) $\{d_1 y_1, d_2 y_2, \dots, d_m y_m\}$ is a basis of K .
- (2) $d_i | d_{i+1}$ for each i .

In particular, K is free of rank $m \leq n$.

Proof. Let $\{x_1, x_2, \dots, x_n\}$ be any basis of W , and let $\pi_i : W \rightarrow R$ be the projection for each i given by $\pi_i(\sum r_k x_k) = r_i$. Define

$$\mathcal{S} = \{\alpha(K) \mid \alpha : W \rightarrow R \text{ is } R\text{-linear}\}.$$

Then \mathcal{S} consists of ideals of R , and \mathcal{S} is nonempty (it contains the zero ideal). Hence, Lemma 2 shows that \mathcal{S} contains a maximal member $\lambda(K)$ where $\lambda : W \rightarrow R$

⁸⁷ Actually this requires a set-theoretical theorem called transfinite recursion. This is discussed in Appendix D.

is R -linear. Observe that $\lambda(K) \neq 0$ (otherwise 0 is maximal in S , so $\pi_i(K) = 0$ for each i , which implies that $K = 0$, contrary to assumption). Since R is a PID, let

$$\lambda(K) = Rd, \quad 0 \neq d \in R.$$

Write $d = \lambda(z)$, where $z \in K$.

Claim 1. $d|\alpha(z)$ for each R -linear $\alpha : W \rightarrow R$.

Proof. Since R is a PID let $e = \gcd(d, \alpha(z))$. Then $e|d$, $e|\alpha(z)$, and $e = rd + s\alpha(z)$ for some $r, s \in R$. With this, define

$$\gamma : W \rightarrow R \text{ by } \gamma(x) = r\lambda(x) + s\alpha(x), \quad \text{for all } x \in W.$$

Then γ is R -linear and $\gamma(z) = rd + s\alpha(z) = e$. Hence, $e \in \gamma(K)$, so $Re \subseteq \gamma(K)$. But $Rd \subseteq Re$ (because $e|d$) so we have $Rd \subseteq Re \subseteq \gamma(K)$. Since $Rd = \lambda(K)$ is maximal in S , it follows that $Rd = Re = \gamma(K)$. In particular, $d = ue$ for some unit $u \in R$, so the fact that $e|\alpha(z)$ implies that $d|\alpha(z)$. This proves Claim 1.

In particular, $d|\pi_i(z)$ for each i by Claim 1, say $\pi_i(z) = c_i d$ where $c_i \in R$. Define $y = \sum_{i=1}^n c_i x_i$. Then $dy = \sum_{i=1}^n dc_i x_i = \sum_{i=1}^n \pi_i(z) x_i = z \in K$ because the π_i are the projections. Hence, $d = \lambda(z) = \lambda(dy) = d\lambda(y)$. Since R is an integral domain, this implies that

$$\lambda(y) = 1.$$

Claim 2. (i) $W = Ry \oplus \ker \lambda$, and (ii) $K = Rdy \oplus (K \cap \ker \lambda)$.

Proof. $Ry \cap \ker \lambda = 0$ because $\lambda(ry) = r\lambda(y) = r$ for each $r \in R$. To see that $W = Ry + \ker \lambda$, let $x \in W$, write it as $x = \lambda(x)y + (x - \lambda(x)y)$, and observe that $\lambda(x - \lambda(x)y) = \lambda(x) - \lambda(x)\lambda(y) = 0$ because $\lambda(y) = 1$. This proves (i).

For (ii), note first that $Rdy \cap (K \cap \ker \lambda) \subseteq Ry \cap \ker \lambda = 0$ by (i). To see that $K = Rdy + (K \cap \ker \lambda)$, let $x_0 \in K$ and observe that $\lambda(x_0)y \in \lambda(K)y = Rdy \subseteq K$. If we write x_0 as $x_0 = \lambda(x_0)y + (x_0 - \lambda(x_0)y)$, then (ii) follows because $\lambda(x_0)y \in Rdy$, $x_0 - \lambda(x_0)y \in K$, and $\lambda(x_0 - \lambda(x_0)y) = \lambda(x_0) - \lambda(x_0)\lambda(y) = 0$. This proves Claim 2.

We can now use these results to complete the proof of Theorem 6 by induction on n . If $n = 1$ then $K \subseteq Rx_1$, and we consider $A = \{a \in R \mid ax_1 \in K\}$. This is an ideal of R , say $A = Rd_1$, $d_1 \in R$. One verifies that $K = Rd_1x_1$, so the bases $\{x_1\}$ and $\{d_1x_1\}$ satisfy our requirements.

Let $n \geq 2$. Then $\ker \lambda$ is free by Lemma 3 and Claim 2(i); and $\ker \lambda$ has rank $n - 1$ by Corollary 3 of Theorem 5 §7.1 because Ry is free of rank 1 (y is torsion-free as $\lambda(y) = 1$). Hence, by induction, there exists a basis $\{y_2, \dots, y_m\}$ of $\ker \lambda$, an integer $m \leq n - 1$, and elements d_2, \dots, d_m in R such that $d_i|d_{i+1}$ for $i \geq 2$ and $\{d_2y_2, \dots, d_my_m\}$ is a basis of $K \cap \ker \lambda$. Hence, $\{y, y_2, \dots, y_m\}$ is a basis of W by Claim 2(i), and $\{dy, d_2y_2, \dots, d_my_m\}$ is a basis of K by Claim 2(ii).

Thus, it remains to show that $d|d_2$. To this end, define $\varphi : W \rightarrow R$ by taking $\varphi(y) = 1 = \varphi(y_2)$ and $\varphi(y_i) = 0$ for $i > 2$. Then $d = \varphi(dy) \in \varphi(K)$, so $Rd \subseteq \varphi(K)$. Since $Rd = \lambda(K)$ is maximal in S , we obtain $Rd = \varphi(K)$. But then we obtain $d_2 = \varphi(d_2y_2) \in \varphi(K) = Rd$, so $d|d_2$ as required. \blacksquare

With this we can prove the main theorem of this chapter.

Theorem 7. Fundamental Theorem. Let R denote a PID. If $_RM$ is finitely generated, there exist integers $m \geq 0$, $k \geq 0$, and nonzero nonunits d_1, d_2, \dots, d_m in R , such that $d_i|d_{i+1}$ for each i and

$$M \cong R/Rd_1 \oplus R/Rd_2 \oplus \cdots \oplus R/Rd_m \oplus R^k.$$

Moreover, $m \in \mathbb{Z}$, $k \in \mathbb{Z}$, and the elements $d_i \in R$ are uniquely determined by M .

Proof. We split the proof into showing that such a decomposition exists, and then that m , k and the d_i are unique.

Existence. If $_RM$ is n -generated, let $\theta : W \rightarrow M$ be an onto R -homomorphism where W is free of rank n . If we write $\ker \theta = K$, the submodule theorem (Theorem 6) provides a basis $\{y_1, y_2, \dots, y_n\}$ of W , an integer $m \leq n$, and nonzero nonunits d_1, d_2, \dots, d_m of R such that $d_i|d_{i+1}$ for each i and $\{d_1y_1, d_2y_2, \dots, d_my_m\}$ is a basis of K . Thus,

$$\begin{aligned} W &= Ry_1 \oplus Ry_2 \oplus \cdots \oplus Ry_m \oplus Ry_{m+1} \oplus \cdots \oplus Ry_n, \\ K &= Rd_1y_1 \oplus Rd_2y_2 \oplus \cdots \oplus Rd_my_m \oplus 0 \oplus \cdots \oplus 0. \end{aligned}$$

Hence, the isomorphism theorem and Corollary 3 of Theorem 3 §7.1 give

$$M \cong W/K \cong \frac{Ry_1}{Rd_1y_1} \oplus \frac{Ry_2}{Rd_2y_2} \oplus \cdots \oplus \frac{Ry_m}{Rd_my_m} \oplus Ry_{m+1} \oplus \cdots \oplus Ry_n$$

Since $Ry_i/Rd_iy_i \cong R/Rd_i$ for each i , and $Ry_j \cong R$ for each j , this proves the existence part of Theorem 7.

Uniqueness. Each factor module R/Rd_i is principal; in fact $R/Rd_i = R(1 + Rd_i)$ and $o(1 + Rd_i) = d_i$. Since $R = R1$ is also principal, we obtain an (equivalent) internal direct sum decomposition of M :

$$M = (Rx_1 \oplus Rx_2 \oplus \cdots \oplus Rx_m) \oplus (Rw_1 \oplus \cdots \oplus Rw_k),$$

where $m \geq 0$, $k \geq 0$, w_j is torsion-free for each j , $o(x_i) = d_i$ for each $i = 1, 2, \dots, m$, and $d_i|d_{i+1}$. Note that this proves Theorem 1 (with an additional uniqueness statement), and hence completes the proofs of Theorems 4 and 5. We use these results to prove uniqueness in the present theorem.

For simplicity, we shall write $T = T(M)$, $X = Rx_1 \oplus Rx_2 \oplus \cdots \oplus Rx_m$, and $W = R w_1 \oplus \cdots \oplus R w_k$, so that $M = X \oplus W$. We have $X \subseteq T$ because $d_i x_i = 0$ for each i , and we claim that this is equality. If $z \in T$ write $z = x + w$ where $x \in X$ and $w \in W$. Then $z - x = w \in T \cap W$, so $z - x$ is both torsion and torsion free (W is free). Hence, $z - x = 0$, so $z = x \in X$ and we have proved that $T \subseteq X$. This shows that $X = T$, and hence that $M = T \oplus W$. But then $W \cong M/T$, and so $k = \text{rank}(M/T)$ is uniquely determined by M .

For the rest, let p_1, p_2, \dots, p_t be the distinct primes dividing at least one of the elements d_1, d_2, \dots, d_m , and write

$$d_i = p_1^{k_{i1}} p_2^{k_{i2}} \cdots p_t^{k_{it}}, \quad k_{ij} \geq 0, \tag{*}$$

where some of the k_{ij} may be zero.

Claim. $Rx_i = Rx_{i1} \oplus Rx_{i2} \oplus \cdots \oplus Rx_{it}$, where $o(x_{ij}) = p_j^{k_{ij}}$ for each j .

Proof. Write $d_{ij} = d_i/p_j^{k_{ij}}$ for each j , and define $x_{ij} = d_{ij}x_i$. Then $o(x_{ij}) = p_j^{k_{ij}}$ because⁸⁸ $o(x_i) = d_{ij}p_j^{k_{ij}}$. Moreover, $d_{i1}, d_{i2}, \dots, d_{it}$ are relatively prime (no prime divides all of them), say $1 = r_1d_{i1} + r_2d_{i2} + \dots + r_td_{it}$ for some $r_i \in R$. It follows that $Rx_i = Rd_{i1}x_i + Rd_{i2}x_i + \dots + Rd_{it}x_i$. This sum is direct by Theorem 4 because Rx_{ij} is contained in the p_j -primary component of Rx_i (in fact $p_j^{k_{ij}}(Rx_{ij}) = 0$). This proves the Claim.

It follows that T is the direct sum of *all* the modules Rx_{ij} , and so its p_j -primary component is the sum of all these summands that are p_j -modules, that is $T(p_j) = Rx_{1j} \oplus Rx_{2j} \oplus \dots \oplus Rx_{mj}$. Thus, the primes p_j are uniquely determined by M (they are the primes p such that $T(p) \neq 0$). Moreover, the fact that $d_i|d_{i+1}$ for each i implies that $k_{1j} \leq k_{2j} \leq \dots \leq k_{mj}$ for each j . Eliminating zero values, this shows that $(k_{mj}, k_{m-1j}, k_{m-2j}, \dots)$ is the type of the p_j -module $T(p_j)$. Hence, the k_{ij} are uniquely determined by M . But then (*) shows that the elements d_1, d_2, \dots, d_m (and hence m) are also uniquely determined. ■

In the proof we obtained a useful internal direct sum decomposition

Corollary 1. *If R is a PID and $_RM$ is finitely generated, there exist $m \geq 0$, $k \geq 0$ in \mathbb{Z} , and nonzero nonunits d_1, d_2, \dots, d_m in R , such that $d_i|d_{i+1}$ for each i and*

$$M = (Rx_1 \oplus Rx_2 \oplus \dots \oplus Rx_m) \oplus (Rw_1 \oplus \dots \oplus Rw_k),$$

where w_j is torsion free for each j and $o(x_i) = d_i$ for each i .

Note that this proves Theorem 1 (which we have used several times).

The elements d_i in Theorem 7 are called the **invariant factors** for the module M . In the notation of the proof, the elements $p_j^{k_{ij}}$ are the elementary divisors of $M(p_j)$ for each j , and so are uniquely determined (up to unit multiples) by the module M . They are called the **elementary divisors** of M .

Specializing Corollary 1 to the case of \mathbb{Z} -modules, we obtain the main motivating example for Theorem 6. We use the fact that every abelian group of order n is isomorphic to \mathbb{Z}_n .

Corollary 2. Fundamental Theorem of Finitely Generated Abelian Groups. *If G is a finitely generated abelian group then*

$$G \cong \mathbb{Z}_{t_1} \oplus \mathbb{Z}_{t_2} \oplus \dots \oplus \mathbb{Z}_{t_m} \oplus \mathbb{Z}^k,$$

where $t_i > 0$ and $t_i|t_{i+1}$ for each i . The integers k, m, t_1, \dots, t_m are uniquely determined by G .

The theorem that every finite abelian group is a direct sum of cyclic groups was first proved in 1870 by Leopold Kronecker. The uniqueness came later as did the extension to finitely generated groups.

We can easily pass back and forth between the primary decomposition of a torsion module and the decomposition in the fundamental theorem in terms of the invariant factors. This is illustrated in Example 8 for a torsion abelian group.

Example 8. Let $G = \mathbb{Z}_{15} \oplus \mathbb{Z}_{90} \oplus \mathbb{Z}_{540}$ so the invariant factors are 15, 90, and 540. Using the primary decomposition theorem on each of these summands, and

⁸⁸In general, if $o(x) = ab \neq 0$ then $o(ax) = b$.

rearranging the resulting summands, we obtain

$$\begin{aligned} G &\cong (\mathbb{Z}_3 \oplus \mathbb{Z}_5) \oplus (\mathbb{Z}_2 \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_5) \oplus (\mathbb{Z}_4 \oplus \mathbb{Z}_{27} \oplus \mathbb{Z}_5) \\ &= (\mathbb{Z}_2 \oplus \mathbb{Z}_4) \oplus (\mathbb{Z}_3 \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_{27}) \oplus (\mathbb{Z}_5 \oplus \mathbb{Z}_5 \oplus \mathbb{Z}_5). \end{aligned}$$

Thus, the primary components (and hence the elementary divisors) are

$$\begin{aligned} G(2) &= \mathbb{Z}_4 \oplus \mathbb{Z}_2 && \text{type } (2, 1), \\ G(3) &= \mathbb{Z}_{27} \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_3 && \text{type } (3, 2, 1), \\ G(5) &= \mathbb{Z}_5 \oplus \mathbb{Z}_5 \oplus \mathbb{Z}_5 && \text{type } (1, 1, 1). \end{aligned}$$

On the other hand, given these primary components, we can retrieve the groups \mathbb{Z}_{540} , \mathbb{Z}_{90} , and \mathbb{Z}_{15} , respectively, as the sum of the summands in the primary components of largest order, second largest order, and so on

$$\begin{aligned} \mathbb{Z}_{540} &= \mathbb{Z}_4 \oplus \mathbb{Z}_{27} \oplus \mathbb{Z}_5, \\ \mathbb{Z}_{90} &= \mathbb{Z}_2 \oplus \mathbb{Z}_9 \oplus \mathbb{Z}_5, \\ \mathbb{Z}_{15} &= 0 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_5. \end{aligned}$$

□

Exercises 7.2

Throughout these exercises R is a PID and modules are R -modules unless otherwise specified.

1. Write down all the abelian groups (up to isomorphism) of each order:

(a) 9	(b) 10	(c) 12	(d) 27
(e) 30	(f) 60	(g) 108	
2. If p is a prime, determine all the abelian groups of order:

(a) p^4	(b) p^6
-----------	-----------
3. If $p \neq q$ are primes, determine all the abelian groups of order:

(a) pq^2	(b) p^2q^2
------------	--------------
4. If p, q , and r are distinct primes, determine how many nonisomorphic abelian groups there are of order:

(a) $p^2q^3r^4$	(b) p^5qr^2
-----------------	---------------
5. List the types of all nonzero subgroups of G if G is a p -group of type $(3, 2, 1)$.
6. If G is an abelian group with $|G| = 108$, and $G(2)$ and $G(3)$ have types (2) and $(2, 1)$, respectively, how many nonisomorphic subgroups does G have?
7. Find the type of the primary components of

(a) $G = \mathbb{Z}_{12} \oplus \mathbb{Z}_{60} \oplus \mathbb{Z}_{75}$	(b) $G = \mathbb{Z}_{36} \oplus \mathbb{Z}_{42} \oplus \mathbb{Z}_{98}$
---	---
8. Determine the abelian groups of order p^n containing

(a) an element of order p^{n-1} ,	(b) an element of order p^{n-2} .
-------------------------------------	-------------------------------------
9. Determine the abelian groups of order p^6 containing

(a) no element of order greater than p^2 ,	(b) no element of order p^4 .
--	---------------------------------
10. Are the groups $\mathbb{Z}_5 \oplus \mathbb{Z}_{10} \oplus \mathbb{Z}_{25} \oplus \mathbb{Z}_{36} \oplus \mathbb{Z}_{54}$ and $\mathbb{Z}_{50} \oplus \mathbb{Z}_{108} \oplus \mathbb{Z}_{450}$ isomorphic?
11. Let $x \in_R M$ have $o(x) = d \neq 0$. If $d = ab$, show that $o(ax) = b$.
12. Let $K \subseteq M$ be modules. Show that

(a) $T(K) = K \cap T(M)$.	(b) If $K \subseteq T(M)$, show that $T(M/K) = T(M)/K$.
----------------------------	---
13. If $K \subseteq M$ are modules, show that M is torsion if and only if both K and M/K are torsion.

14. Show that \mathbb{Q}/\mathbb{Z} is a torsion group that is not finite.
15. If $M = M_1 \oplus \cdots \oplus M_n$ are modules, show that $T(M) = T(M_1) \oplus \cdots \oplus T(M_n)$.
16. Let $K \subseteq M$ be finitely generated abelian groups. Show that
- $K/T(K)$ is isomorphic to a subgroup of $M/T(M)$.
 - $T(K) \subseteq T(M)$ and $T(M)/T(K)$ is isomorphic to a subgroup of $T(M/K)$.
17. Let M be an abelian group and assume that $M = H \oplus W$, where H is torsion and W is torsion free. Show that $H = T(M)$ and $W \cong M/T(M)$.
18. If $\alpha : M \rightarrow N$ is a homomorphism of R -modules, show that $\alpha[T(M)] \subseteq T(N)$, and that there is a unique homomorphism $\bar{\alpha} : M/T(M) \rightarrow N/T(N)$ satisfying $\bar{\alpha}\varphi = \theta\alpha$, where $\varphi : M \rightarrow M/T(M)$ and $\theta : H \rightarrow N/T(N)$ are the coset maps.
19. Describe $T(\mathbb{C}^*)$ and $T(\mathbb{Q}^*)$.
20. Let $d \in R$. If N is any module define $L_d(N) = \{x \in N \mid dx = 0\}$.
- Show that $L_d(N)$ is a submodule of N .
 - If $d \neq 0$, show that $L_d(R/Rd^k) = R(d^{k-1} + Rd^k)$ for all $k \geq 1$.
 - If $M = \bigoplus_{i=1}^m M_i$, show that $L_d(M) = \bigoplus_{i=1}^m L_d(M_i)$.
21. Let $d \in R$. If N is any module define $dN = \{dx \mid x \in N\}$.
- Show that dN is a submodule of N .
 - If $M = \bigoplus_{i=1}^m M_i$, show that $dM = \bigoplus_{i=1}^m dM_i$.
22. Given a prime $p \in R$ and a module $_RM$, define $L_p(M) = \{x \in M \mid px = 0\}$ and $pM = \{px \mid x \in M\}$. (These are submodules and preserve direct sums by the preceding exercises.)
- If $M = Rx$ where $o(x) = p^m$, $m \geq 1$, show that $L_p(M) = R(p^{m-1} + Rp^k)$ and $p(M) = \begin{cases} Rpx & \text{if } m > 1 \\ 0 & \text{if } m = 1 \end{cases}$.
 - If M is a p -module of type (m_1, m_2, \dots, m_t) , show that $L_p(M)$ has type $(1, 1, \dots, 1)$ with t ones. Also show that, if $pM \neq 0$, it has type $(m_1 - 1, m_2 - 1, \dots, m_s - 1)$, where $m_s > 1$ but $m_{s+1} = \dots = m_t = 1$.
23. If p is a prime, determine the structure of a finite abelian group G if $pg = 0$ for all $g \in G$.
24. Show that every submodule of a finitely generated module over a PID is again finitely generated. [Hint: Submodule Theorem.]
25. If G is abelian and n divides $|G|$, show that G has a subgroup of order n . Show that this conclusion fails if G is not abelian.
26. Let $G = \mathbb{Z}x$ where $o(x) = p^n$, p a prime. If $k \leq n$ show that $L_{p^k}(G) = \mathbb{Z}p^{n-k}x$, and hence that $|L_{p^k}(G)| = p^k$.
27. Let G be a finite abelian p -group of type (n_1, n_2, \dots, n_r) .
- Show that $|L_p(G)| = p^r$, so there are $p^r - 1$ elements of order p .
 - Let s be the number of integers i such that $n_i \geq 2$ (possibly $s = 0$). Show that $|L_{p^2}(G)| = p^{r+s}$, and hence that G has $p^r(p^s - 1)$ elements of order p^2 .
28. If $M \oplus M \cong N \oplus N$, where M and N are finitely generated p -modules, show that $M \cong N$.
29. If $G \oplus H \cong G \oplus K$, where G, H , and K are finite abelian p -groups, show that $H \cong K$ —the cancellation property. [Hint: Reduce to the case where G is cyclic.]

$$\begin{array}{ccc} M & \xrightarrow{\alpha} & N \\ \varphi \downarrow & & \downarrow \theta \\ M/T(M) & \xrightarrow{\bar{\alpha}} & N/T(N) \end{array}$$

Chapter 8

p-Groups and the Sylow Theorems

Mathematics is the tool specially suited for dealing with abstract concepts of any kind.
There is no limit to its power in this field.

—Paul Adrien Maurice Dirac

Historically, the theory of groups was concerned only with groups of permutations of a set. This point of view is reinforced by Cayley's theorem, which shows that every abstract group can be viewed as a subgroup of a group of permutations. The concept of an abstract group became important because it focuses attention on those aspects of a group of permutations that do not depend on the underlying set. However, this abstract formulation of the theory loses sight of the combinatorial aspects that are more in evidence for groups of permutations. And these *counting* methods give important information about abstract groups. The best example is Lagrange's theorem, which is based on the fact that a subgroup partitions the group into cosets each having the same number of elements as the subgroup.

In Section 8.2, we derive another such counting theorem, the class equation, from a partition of a finite group and use it, among other things, to deduce many properties of groups of prime power order. Then, in Section 8.3, we present a far-reaching counting method that includes the proof of Lagrange's theorem and the class equation and which, in Section 8.4, we use to prove the Sylow theorems. These beautiful results guarantee the presence of subgroups of prime power order in every finite group and inform us about how many such subgroups there are.

8.1 PRODUCTS AND FACTORS

An important goal of the theory of groups is to prove *structure theorems*, that is to describe all groups of a certain type in terms of particular constructions of well-known subgroups of that type. For example, in Section 7.2 we showed that every finite abelian group is a finite direct product of cyclic subgroups. We begin this section by describing when a group G is isomorphic to a finite direct product $G_1 \times G_2 \times \cdots \times G_n$ of subgroups of G . These direct factors G_i are also images of G , and we gain some insight as to how to describe these images in the second part of this section.

Products of Subgroups

If X and Y are nonempty subsets of a group, define their **product** as follows:

$$XY = \{xy \mid x \in X \text{ and } y \in Y\}.$$

This is an associative multiplication; indeed if Z is another nonempty subset then

$$(XY)Z = \{xyz \mid x \in X, y \in Y \text{ and } z \in Z\} = X(YZ).$$

as the reader can verify. Moreover, $\{1\}X = X = X\{1\}$ for all nonempty sets X , so the set of all nonempty subsets of G is a monoid with unity $\{1\}$. Moreover, the map $a \mapsto \{a\}$ is a group embedding (one-to-one homomorphism), so we identify G as a subgroup of this monoid by identifying $a = \{a\}$ for all $a \in G$. Hence, we write $X\{a\} = Xa$ and $\{a\}X = aX$, which agrees with our earlier usage for cosets Ha and conjugates $a^{-1}Ha$ of a subgroup H .

These products are most useful for subgroups. If H and K are subgroups of some group, the product HK is again a subgroup if and only if $HK = KH$ (Lemma 2 §2.8), and this holds if either $H \triangleleft G$ or $K \triangleleft G$ (where $K \triangleleft G$ means K is normal in G). The following result will be used several times in this chapter.

Lemma 1. Modular law. *If H , K , and M are subgroups of a group and $H \subseteq M$ then $H(K \cap M) = HK \cap M$.*

Proof. The inclusion $H(K \cap M) \subseteq HK \cap M$ is clear because $H \subseteq M$. If $x \in HK \cap M$, say $x = hk = m$, then $k = h^{-1}m \in K \cap M$, so $x = hk \in H(K \cap M)$. \square

For reference, Theorem 6 §2.8 reads

$$\text{If } H \triangleleft G, K \triangleleft G, \text{ and } H \cap K = \{1\} \text{ then } HK \cong H \times K. \quad (*)$$

The next theorem is an important extension of this which requires only that $K \triangleleft G$. Recall that the isomorphism theorem for groups asserts that, if $\alpha : G \rightarrow H$ is a group homomorphism, then $\ker \alpha \triangleleft G$ and $G/\ker \alpha \cong G\alpha$.

Theorem 1. Second Isomorphism Theorem. *Let H and K be subgroups of a group G with $K \triangleleft G$. Then HK is a subgroup of G , $K \triangleleft HK$, $H \cap K \triangleleft H$, and*

$$\frac{HK}{K} \cong \frac{H}{H \cap K}.$$

Proof. HK is a subgroup by Lemma 2 §2.8, and $K \triangleleft HK$ because $K \subseteq HK$. Define $\alpha : H \rightarrow HK/K$ by $\alpha(h) = Kh$. (Note that Kh is in HK/K because $H \subseteq HK$.) Then α is a homomorphism, and it is onto because, if $x \in HK = KH$, say $x = kh$,

then $Kx = K(kh) = (Kh)h = Kh = \alpha(h)$. Now the theorem follows from the isomorphism theorem because $\ker(\alpha) = \{h \in H \mid Kh = K\} = H \cap K$. ■

If H and K are both finite subgroups, we saw in Corollary 1 to Theorem 6 §2.8 that $|HK| = |H||K|$ whenever $H \cap K = \{1\}$. Here is a useful generalization.

Theorem 2. *Let H and K be finite subgroups of a group. Then*

$$|HK||H \cap K| = |H||K|.$$

Proof. Write $N = H \cap K$ for convenience. Let Nk_1, \dots, Nk_m be the distinct cosets of N in K , so that $m = |K : N|$ is the index of N in K .

Claim. $HK = Hk_1 \cup Hk_2 \cup \dots \cup Hk_m$, a disjoint union.

Proof. If $k \in K$, then $k = nk_i$ for some $n \in N$. If $h \in H$ then $hk = (hn)k_i \in Hk_i$, which proves that $HK \subseteq Hk_1 \cup \dots \cup Hk_m$. Thus, $HK = Hk_1 \cup \dots \cup Hk_m$. To see that this is a disjoint union, suppose that $Hk_i \cap Hk_j$ is nonempty. Then $Hk_i = Hk_j$, so $k_ik_j^{-1} \in H \cap K = N$. Hence, $Nk_i = Nk_j$, so $i = j$. This proves the Claim.

Finally, the Claim gives $|HK| = m|H| = |K : N||H| = \frac{|K|}{|N|}|H|$, as required. ■

We now give two generalizations of Theorem 6 §2.8—see (*) above; the first drops the condition that $H \cap K = \{1\}$.

Theorem 3. *If $H \triangleleft G$ and $K \triangleleft G$ then $\frac{HK}{H \cap K} \cong \frac{H}{H \cap K} \times \frac{K}{H \cap K}$.*

Proof. Write $N = H \cap K$ for convenience, and define

$$\alpha : HK \rightarrow \frac{H}{N} \times \frac{K}{N} \quad \text{by} \quad \alpha(hk) = (Nh, Nk).$$

Then α is well defined because $hk = h_1k_1$ means $h_1^{-1}h = k_1k^{-1} \in H \cap K = N$. Hence, $hN = h_1N$, so $Nh = Nh_1$ because $N \triangleleft G$. Similarly, $Nk = Nk_1$.

Claim. α is a homomorphism.

Proof. Write $x = hk$ and $y = h_1k_1$ in $HK = KH$, and write $kh_1 = h_2k_2$ in HK . Then $xy = hh_2k_2k_1$ so $\alpha(xy) = (NhNh_2, Nk_2Nk_1)$. On the other hand, we have $\alpha(x)\alpha(y) = (NhNh_1, NkNk_1)$, so we must prove that $NhNh_2 = NhNh_1$ and $Nk_2Nk_1 = NkNk_1$, equivalently (since G/N is a group) $Nh_2 = Nh_1$ and $Nk_2 = Nk$. But we have $kh_1 = h_2k_2$, so

$$h_1h_2^{-1} = k^{-1}(h_2k_2h_2^{-1}) \in H \cap K = N \text{ because } K \triangleleft G, \text{ and}$$

$$k_2k^{-1} = h_2^{-1}(kh_1k^{-1}) \in K \cap H = N \text{ because } H \triangleleft G.$$

Hence, $Nh_2 = Nh_1$ and $Nk_2 = Nk$, which proves the Claim.

With the Claim, α is an onto homomorphism, so we are done by the isomorphism theorem because $\ker \alpha = \{hk \mid Nh = N \text{ and } Nk = N\} = N$. ■

Our second generalization of Theorem 6 §2.8—see (*) above; is to extend it to more than two factors. We require the following result, interesting in itself.

Lemma 2. *The following are equivalent for subgroups G_1, G_2, \dots, G_n of a group:*

- (1) $(G_1G_2 \dots G_{k-1}) \cap G_k = \{1\}$ for each $k = 2, 3, \dots, n$.
- (2) If $g_1g_2 \dots g_n = 1$, where $g_i \in G_i$ for each i , then $g_i = 1$ for each i .

Proof. Given (1), if $g_1g_2 \cdots g_n = 1$ then $g_n \in (G_1G_2 \cdots G_{n-1}) \cap G_n = \{1\}$, so $g_n = 1$ and $g_1g_2 \cdots g_{n-1} = 1$. Now repeat the procedure to obtain $g_{n-1} = 1$. Continue to prove (1) \Rightarrow (2). The proof that (2) \Rightarrow (1) is left to the reader. ■

Call subgroups G_1, G_2, \dots, G_n **unconnected** if the conditions in Lemma 2 are satisfied. Using this notion, we now give a useful characterization of when a group is isomorphic to a direct product of finitely many subgroups.

Theorem 4. Let G be a group and assume that $G = G_1G_2 \cdots G_n$ where the G_i are subgroups. Then the following conditions are equivalent:

- (1) The G_i are unconnected and $G_k \triangleleft G$ for each k .
- (2) The G_i are unconnected and $g_i g_j = g_j g_i$ when $g_i \in G_i$, $g_j \in G_j$ and $i \neq j$.
- (3) $(G_1 \cdots G_{k-1} G_{k+1} \cdots G_n) \cap G_k = \{1\}$ and $G_k \triangleleft G$ for each k .
- (4) $(G_1 \cdots G_{k-1} G_{k+1} \cdots G_n) \cap G_k = \{1\}$ for each k and $g_i g_j = g_j g_i$ whenever $g_i \in G_i$, $g_j \in G_j$ and $i \neq j$.

In this case $G \cong G_1 \times G_2 \times \cdots \times G_n$, and each $g \in G$ is uniquely represented as a product $g = g_1g_2 \cdots g_n$, where $g_i \in G_i$ for each i .

Proof. (1) \Rightarrow (2). Given (1), let $g_i \in G_i$, $g_j \in G_j$ and $i \neq j$. Assume $i < j$, so that $G_i \cap G_j \subseteq (G_1G_2 \cdots G_{j-1}) \cap G_j = \{1\}$. If we write $a = g_i g_j g_i^{-1} g_j^{-1}$, then $a \in G_i$ because $g_j g_i^{-1} g_j^{-1} \in G_i \triangleleft G$. Similarly $a \in G_j$, so $a \in G_i \cap G_j = \{1\}$. Hence $a = 1$, so $g_i g_j = g_j g_i$, as required.

(2) \Rightarrow (1). Let $b \in G_k$, $a \in G$; we want $a^{-1}ba \in G_k$. If $a = a_1a_2 \cdots a_n$, $a_i \in G_i$ then, since b commutes with each of $a_1^{-1}, a_2^{-1}, \dots, a_{k-1}^{-1}$ by (2), we have

$$a^{-1}ba = (a_n^{-1} \cdots a_k^{-1})b(a_k \cdots a_n) = a_n^{-1} \cdots a_{k+1}^{-1}(a_k^{-1}ba_k)a_{k+1} \cdots a_n.$$

But $a_k^{-1}ba_k \in G_k$ and so commutes with each of a_{k+1}, \dots, a_n . It follows that $a^{-1}ba = a_k^{-1}ba_k \in G_k$, so $G_k \triangleleft G$. This proves (1).

(3) \Leftrightarrow (4) and (4) \Rightarrow (2). Both (3) and (4) imply that the G_i are unconnected.

(2) \Rightarrow (4). Choose $b_k \in G_k \cap (G_1 \cdots G_{k-1} G_{k+1} \cdots G_n)$ and write $b_k = a_1 \cdots a_{k-1} \cdot a_{k+1} \cdots a_n$ with $a_i \in G_i$ for each i . Since b_k commutes with these a_i , this gives $1 = a_1 a_{k-1} b_k^{-1} a_{k+1} \cdots a_n$, so $b_k = 1$ by (2) and Lemma 2. This proves (4).

Now define $\theta : G_1 \times G_2 \times \cdots \times G_n \rightarrow G$ by $\theta(g_1, g_2, \dots, g_n) = g_1g_2 \cdots g_n$. This is onto because $G = G_1G_2 \cdots G_n$, and it is a homomorphism because the G_i commute elementwise. (Note, this does not require that the G_i are abelian.) As θ is one-to-one (because the G_i are unconnected), θ is an isomorphism, and so $G \cong G_1 \times G_2 \times \cdots \times G_n$. Finally, if $g = a_1a_2 \cdots a_n = b_1b_2 \cdots b_n$, $a_i \in G_i$, $b_i \in G_i$; we must show $a_i = b_i$ for all i . First $(b_1^{-1}a_1)a_2a_3 \cdots a_n = b_2b_3 \cdots b_n$. Since b_2 commutes with $(b_1^{-1}a_1)$ by (2), we get $(b_1^{-1}a_1)(b_2^{-1}a_2)a_3 \cdots a_n = b_3 \cdots b_n$. Continue to get $(b_1^{-1}a_1)(b_2^{-1}a_2) \cdots (b_n^{-1}a_n) = 1$. Hence, each $b_i^{-1}a_i = 1$ by Lemma 2, proving the last sentence of the Theorem. ■

The Correspondence Theorem

If $\alpha : G \rightarrow H$ is an onto group homomorphism, the group $H = \alpha(G)$ is called a **homomorphic image** of G , or simply an **image** of G . Thus, the image $\alpha(G)$ enjoys any property of G that is preserved by homomorphisms, for example, being abelian

or cyclic. The image is a simplified version of the group and so is easier to study. The idea is to learn about the group G by investigating its homomorphic images.

The isomorphism theorem (Theorem 4 §2.10) provides a fundamental tool for investigating a homomorphic image $\alpha(G)$ of a group G . It asserts that $\alpha(G)$ is isomorphic to a factor group of G , indeed that $\alpha(G) \cong G/K$, where K denotes the kernel of α . On the other hand, if K is any normal subgroup of G , the factor group G/K is a homomorphic image of G via the coset homomorphism $G \rightarrow G/K$ given by $g \mapsto Kg$. Thus, studying the images $\alpha(G)$ of G is the same as studying the factors G/K of G . The factors of G are very closely connected to G itself and we now focus on these factors and how to use them to study the group G .

Because many properties of G can be described in terms of the subgroups of G , we need to be able to obtain information about these subgroups from knowledge of the subgroups of a factor group G/K . The next theorem provides a method of doing this. It gives a very useful correspondence between the set of subgroups of G that contain K and the set of all subgroups of G/K . Moreover, the correspondence is such that, if we can determine all the subgroups in one of these sets, we can easily compute the subgroups in the other set.

To show how this correspondence works let $K \triangleleft G$, and consider the set of subgroups H of G such that $K \subseteq H$. For any such H , define

$$H/K = \{Kh \mid h \in H\}.$$

Lemma 3. If $K \subseteq H \subseteq G$ are groups and $K \triangleleft G$, then

- (1) H/K is a subgroup of G/K .
- (2) If $K \subseteq H_1 \subseteq G$, then $H \subseteq H_1$ if and only if $H/K \subseteq H_1/K$.

Proof. (1) This is because H/K is the image of H under the coset map $G \rightarrow G/K$.

(2) Clearly $H \subseteq H_1$ implies $H/K \subseteq H_1/K$. Conversely, assume $H/K \subseteq H_1/K$, and let $h \in H$. Then $Kh \in H/K = H_1/K$, say $Kh = Kh_1$ for some $h_1 \in H_1$. As $K \subseteq H_1$ this gives $h \in Kh_1 \subseteq H_1h_1 \subseteq H_1$, so $H \subseteq H_1$. \square

Theorem 5. Correspondence Theorem. If $K \triangleleft G$ are groups, define a map

$$\Theta : \{H \mid K \subseteq H \subseteq G, H \text{ is a subgroup}\} \rightarrow \{\mathcal{H} \mid \mathcal{H} \subseteq G/K \text{ is a subgroup}\}$$

by $\Theta(H) = H/K$. Let H , H_1 and H_2 be subgroups of G containing K . Then:

- (1) The map Θ is a bijection.
- (2) If $\mathcal{H} \subseteq G/K$ is a subgroup, then $\mathcal{H} = H/K$ where $H = \{g \in G \mid Kg \in \mathcal{H}\}$.
- (3) Θ preserves containment: $H \subseteq H_1$ if and only if $H/K \subseteq H_1/K$.
- (4) Θ preserves normality: $H \triangleleft H_1$ if and only if $H/K \triangleleft H_1/K$.
- (5) Θ preserves intersections: $(H_1/K) \cap (H_2/K) = (H_1 \cap H_2)/K$.
- (6) Θ preserves products: $(H_1/K)(H_2/K) = (H_1H_2)/K$ if H_1H_2 is a subgroup.

Proof. (1) The map Θ is one-to-one by Lemma 2, and it is onto by (2).

(2) Given a subgroup $\mathcal{H} \subseteq G/K$, define $H = \{h \in G \mid Kh \in \mathcal{H}\}$ as in (2). Then H is a subgroup of G and $K \subseteq H$. Moreover $\mathcal{H} \subseteq H/K$ because $Kg \in \mathcal{H}$ implies that $g \in H$, whence $Kg \in H/K$. Conversely, if $Kg \in H/K$ then $Kg = Kh$ for some $h \in H$ so, as $K \subseteq H$, $g \in Kh \subseteq Hh \subseteq H$. But then $Kg \in \mathcal{H}$ by the definition of H , and we have shown that $H/K \subseteq \mathcal{H}$. Hence $H/K = \mathcal{H}$, as required.

(3) This restates Lemma 2.

(4) Assume $H/K \triangleleft H_1/K$. To show $H \triangleleft H_1$, let $h \in H$ and $h_1 \in H_1$. Then

$$K(h_1hh_1^{-1}) = Kh_1Kh(Kh_1)^{-1} \in Kh_1(H/K)(Kh_1)^{-1} \subseteq H/K.$$

Hence $K(h_1hh_1^{-1}) = Kh'$ for some $h' \in H$, so $h_1hh_1^{-1} \in KH \subseteq H$ because $K \subseteq H$. As $h \in H$ was arbitrary, this shows that $H \triangleleft H_1$.

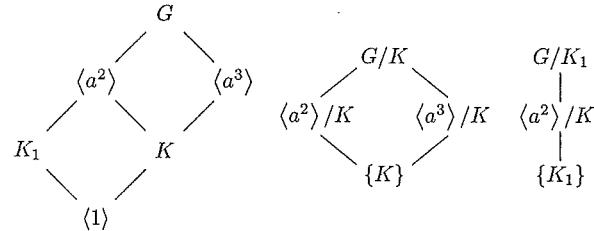
Conversely, let $H \triangleleft H_1$. Then $(Kh_1)^{-1}KhKh_1 = K(h_1^{-1}hh_1) \in H/K$ for all $h_1 \in H_1$, $h \in H$, because $h_1^{-1}hh_1 \in H$. As $h \in H$ was arbitrary, this shows that $(Kh)^{-1}(H/K)Kh \subseteq H/K$. Hence $H/K \triangleleft H_1/K$, as required. \blacksquare

(5) and (6) These are left as Exercises 9 and 10. \blacksquare

If $K \triangleleft G$, the bijection $H \leftrightarrow H/K$ in Theorem 5 pairs all subgroups $H \supseteq K$ of G with all subgroups H/K of G/K . Not only is this correspondence a bijection, it also pairs normal subgroups with normal subgroups by (4) and preserves inclusion by (3). This last fact means that the lattice diagram of all subgroups of G/K has the same form as the lattice diagram of all subgroups of G that contain K . In particular, the bijection pairs G with G/K , and it pairs K with $K/K = \{K\}$ —the trivial subgroup of G/K . This is illustrated in the following two examples.

Example 1. Let $G = \langle a \rangle$, where $o(a) = 12$, and let $K = \langle a^6 \rangle$ and $K_1 = \langle a^4 \rangle$. Draw the lattice diagram of all subgroups of G , and use the correspondence theorem to obtain the lattice of all subgroups of G/K and G/K_1 .

Solution. The subgroups of G are given by the divisors of 12, and the subgroup lattice for G appears on the left in the diagram (see Example 14 §2.4). The subgroups of G/K are thus determined (using the correspondence theorem) by the subgroups G , $\langle a^3 \rangle$, $\langle a^2 \rangle$, and K of G that contain K . Thus the lattice diagram for G/K shown in the center diagram can be “read off” from the diagram for G .



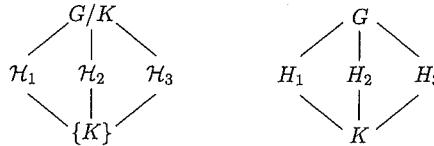
Similarly G , $\langle a^2 \rangle$ and K_1 are the only subgroups containing K_1 , and they give the subgroup lattice for G/K_1 on the right. \square

Example 2. Consider the octic group $G = D_4 = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$, where $o(a) = 4$, $o(b) = 2$ and $aba = b$. If $K = \{1, a^2\}$, determine all the subgroups H of G such that $K \subseteq H \subseteq G$.

Solution. By Example 4 §2.9, $K = Z(G)$, and $G/K = \{K, Ka, Kb, Kba\} \cong K_4$, the Klein group. Hence, the subgroups of G/K are $\{K\}$, G/K , and

$$\mathcal{H}_1 = \langle Ka \rangle = \{K, Ka\}; \quad \mathcal{H}_2 = \langle Kb \rangle = \{K, Kb\}; \quad \mathcal{H}_3 = \langle Kba \rangle = \{K, Kba\}.$$

The subgroup lattice diagram of G/K is shown at the left in the diagram.



The correspondence theorem ensures that, for each i , there is a unique subgroup H_i of G such that $K \subseteq H_i$ and $\mathcal{H}_i = H_i/K$. Explicitly, (2) of Theorem 5 gives

$$\begin{aligned} H_1 &= \{g \in G \mid Kg \in \mathcal{H}_1\} = \{1, a, a^2, a^3\}, \\ H_2 &= \{g \in G \mid Kg \in \mathcal{H}_2\} = \{1, b, a^2, ba^2\}, \\ H_3 &= \{g \in G \mid Kg \in \mathcal{H}_3\} = \{1, ba, a^2, ba^3\}. \end{aligned}$$

The correspondence theorem also shows that these are the only subgroups H such that $K \subseteq H \subseteq G$, and that the lattice of such subgroups (in the right diagram) has the same form as the subgroup lattice of G/K . Furthermore, the fact that \mathcal{H}_1 , \mathcal{H}_2 , and \mathcal{H}_3 are normal in (the abelian group) G/K guarantees that H_1 , H_2 , and H_3 are normal in G . (Of course this also follows because they are of index 2 in G). \square

An important special case of the correspondence theorem describes when a factor group is simple. If $K \triangleleft G$, the group G/K is simple if and only if the only normal subgroups are the trivial subgroup $K/K = \{K\}$ and the whole group G/K . Hence, the correspondence theorem shows that the only normal subgroups H such that $K \subseteq H \subseteq G$ are $H = K$ and $H = G$. A normal subgroup $K \neq G$ with this latter property is called a **maximal normal subgroup** of G . This discussion is summarized in

Theorem 6. *A normal subgroup $K \triangleleft G$ is a maximal normal subgroup if and only if G/K is simple.*

Every finite group $G \neq \{1\}$ has maximal normal subgroups—choose any proper normal subgroup (possibly $\{1\}$) of maximal order. Hence, G has finite simple factor groups by Theorem 2, which shows that finite simple groups are quite common. In fact they serve as “building blocks” by which we can study the structure of finite groups in general. We return to this topic in Chapter 9.

The correspondence theorem describes the subgroups of a factor group G/K , where $K \triangleleft G$. We now turn to the *images* of G/K . Of course they are all images of G , and so have the form G/H for some $H \triangleleft G$. They are described next by an important consequence of the isomorphism theorem that will be needed later.

Theorem 7. Third Isomorphism Theorem. *Let $K \subseteq H \subseteq G$ be groups, where $K \triangleleft G$ and $H \triangleleft G$. Then $H/K \triangleleft G/K$ and*

$$\frac{G/K}{H/K} \cong G/H.$$

Proof. Define $\alpha : G/K \rightarrow G/H$ by $\alpha(Kg) = Hg$ for all g in G . This is well defined because $Kg = Kg_1$ implies $gg_1^{-1} \in K \subseteq H$, whence $Hg = Hg_1$. With this it is easy to verify that α is an onto homomorphism, and $\ker(\alpha) = \{Kg \mid Hg = H\} = H/K$. Hence, the isomorphism theorem (Theorem 4 §2.10) completes the proof. \blacksquare

Our final example provides a good illustration of how the second and third isomorphism theorems are used. A group G is called a **metacyclic group** if a normal subgroup $K \triangleleft G$ exists such that both K and G/K are cyclic. Every cyclic

group is metacyclic (take $K = \{1\}$) as is D_n (take K to be the cyclic subgroup of index 2). Thus, D_n is metacyclic but not cyclic.

Example 3. Show that every subgroup and every image of a metacyclic group is again metacyclic.

Solution. Let G be metacyclic, say K and G/K are both cyclic where $K \triangleleft G$.

If H is a subgroup of G then $H \cap K \triangleleft H$, and $H \cap K$ is cyclic (being a subgroup of the cyclic group K). On the other hand, $H/(H \cap K) \cong HK/K$ by the second isomorphism theorem, and HK/K is cyclic (it is a subgroup of G/K). It follows that $H/(H \cap K)$ is cyclic, whence H is metacyclic.

Now let G/N be any image of G , where $N \triangleleft G$. Then $NK \triangleleft G$ by Theorem 1, so $NK/N \triangleleft G/N$. Moreover, NK/N is cyclic because $NK/N \cong K/(N \cap K)$ is an image of the cyclic group K . On the other hand, $(G/N)/(NK/N) \cong G/NK$ is cyclic because $G/NK \cong (G/K)/(NK/K)$ is an image of the cyclic group G/K . This means that G/N is metacyclic. \square

Exercises 8.1

1. Let $S_3 = \{\epsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\}$ where $\sigma^3 = \epsilon = \tau^2$ and $\sigma\tau\sigma = \tau$. Compute XY if:
 - (a) $X = \{\tau, \tau\sigma\}$ and $Y = \{\tau, \tau\sigma^2\}$.
 - (b) $X = \{\sigma, \tau\sigma\}$ and $Y = \{\sigma, \sigma^2\}$.
2. If $\alpha : G \rightarrow C_6$ is an onto group homomorphism and $|\ker(\alpha)| = 3$, show that $|G| = 18$ and G has normal subgroups of orders 3, 6 and 9.
3. Use the correspondence theorem to show that each subgroup H of G with $G' \subseteq H$ is normal in G . (See Theorem 3 §2.9.)
4. In each case use Theorem 5 to find all subgroups of G that contain K .
 - (a) $G = D_6$ and $K = Z(D_6)$.
 - (b) $G = Q$ and $K = Z(Q)$.
 - (c) $G = A_4$ and $K = \{\epsilon, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$.
5. In each case describe all maximal normal subgroups of G .
 - (a) $G = \mathbb{Z}$
 - (b) G is cyclic, $|G| = n$
 - (c) $G = D_{10}$
 - (d) $G = Q$
6. Let $K \triangleleft G$ be such that both K and G/K are simple. Show that either K is the only proper normal subgroup of G , or $G \cong K \times (G/K)$.
7. Let $K \triangleleft G$ and assume that G/K is cyclic, $|K| = k$, and $|G| = n$. If m is an integer such that $k|m$ and $m|n$, show that there is a unique subgroup H such that $K \subseteq H \subseteq G$ and $|H| = m$. [Hint: Theorem 9 §2.4.]
8. If $K \triangleleft G$, show that the following conditions are equivalent.
 - (1) The only subgroups H such that $K \subseteq H \subseteq G$ are $H = K$ and $H = G$.
 - (2) G/K is cyclic and of prime order.
9. Show that the correspondence theorem preserves intersections. More precisely, if $K \subseteq H, H_1 \subseteq G, H, H_1$ subgroups, $K \triangleleft G$, show that $(H/K) \cap (H_1/K) = (H \cap H_1)/K$.
10. Show that the correspondence theorem preserves products. More precisely, if we have $K \subseteq H \subseteq G$ and $K \subseteq H_1 \subseteq G$ where $K \triangleleft G$, H, H_1 subgroups, show that
 - (a) $(H/K) \cdot (H_1/K)$ is a subgroup of G/K if and only if HH_1 is a subgroup of G .
 - (b) In this case $(H/K) \cdot (H_1/K) = (HH_1)/K$.
11. If X and Y are nonempty subsets of some group, show that $\langle X \rangle \langle Y \rangle \subseteq \langle X \cup Y \rangle$, with equality if and only if $\langle X \rangle \langle Y \rangle = \langle Y \rangle \langle X \rangle$. [Hint: Lemma 2 §2.8.]

12. (a) If H is a subgroup of a group G , show that $H^2 = H$.
 (b) If $X \neq \emptyset$ is a finite subset of G , show that X is a subgroup if and only if $X^2 \subseteq X$.
13. Let G be a group with $|G| = pqr$, where p, q and r are distinct primes. If H and K are subgroups, $|H| = pq$, and $|K| = qr$, show that $|H \cap K| = q$. [Hint: Lagrange's theorem.]
14. Let $G = \langle g \rangle$ be a cyclic group, and let $A = \langle g^a \rangle$ and $B = \langle g^b \rangle$. Show that $AB = \langle g^d \rangle$ where $d = \gcd(a, b)$.
15. Let K, A and B be subgroups of G with $K \triangleleft G$ and $A \triangleleft B$. Show that $KA \triangleleft KB$.
16. Let H, K and M be subgroups of a group G , and assume that $H \subseteq M$. If both $H \cap K = M \cap K$ and $HK = MK$ hold, show that $H = M$. [Hint: First show that $M = (HK) \cap M$.]
17. Let $|G| = p^n m$ where p is a prime and p does not divide m . If $K \triangleleft G$ satisfies $|K| = p^n$, show that K is the only subgroup of G of order p^n . [Hint: Theorem 1.]
18. If G is a group, $M \triangleleft G$ is maximal normal, $K \triangleleft G$, and $K \not\subseteq M$, show that
 (a) $G = KM$.
 (b) $K/(K \cap M)$ is simple.
 (c) $G/(K \cap M)$ has a simple direct factor.
19. A group G is called a **metabelian** group if $K \triangleleft G$ exists such that both K and G/K are abelian.
 (a) Show that every subgroup and factor group of a metabelian group is metabelian.
 (b) Show that G is metabelian if and only if the commutator subgroup G' is abelian.
20. Let \mathcal{C} be a nonempty family of groups closed under taking subgroups and images (for example the abelian or the cyclic groups). Call a group G **meta- \mathcal{C}** if there exists $K \triangleleft G$ such that both K and G/K are in \mathcal{C} . Note that every group in \mathcal{C} is meta- \mathcal{C} because the trivial group $\{1\}$ is in \mathcal{C} . If G is meta- \mathcal{C} show that every subgroup and factor group of G is meta- \mathcal{C} .
21. Let G be a group with subgroups H and K . Assume that $|H| = pq$ and $|K| = q^2$, where $p \neq q$ are primes. If $|G| < pq^3$, show that $|H \cap K| = q$.
22. Let G be a finite abelian group.
 (a) If G has two distinct elements of order 2, show that 4 divides $|G|$.
 (b) If G has three distinct elements of order 3, show that 9 divides $|G|$.
23. If G is a group, let M denote the monoid of nonempty subsets of G , and identify $G \subseteq M$ by writing $g = \{g\}$ for each $g \in G$. Show that G is the group of units of M .
24. Let G be a group, let S_G be the group of permutations of G , and write $A = \text{aut}(G)$. If $\tau_a : G \rightarrow G$ is defined by $\tau_a(g) = ag$ for all $g \in G$, let $\tilde{G} = \{\tau_a \mid a \in G\}$ be the group of **translations**. Thus $G \cong \tilde{G}$ by Cayley's theorem (Theorem 6 §2.5).
 (a) Show that $\tilde{G}A$ is a subgroup of S_G called the **holomorph** of G .
 (b) Show that $\tilde{G} \cap A = \{1_G\}$.
 (c) Show that $\tilde{G} \triangleleft \tilde{G}A$.
 (d) Show that $\tilde{G}A/\tilde{G} \cong A$.

8.2 CAUCHY'S THEOREM

If p is a prime and G is a group of order p^n , every element of G has order a power of p by Lagrange's theorem. The converse is also true. If every element of a finite group G has p -power order, then $|G| = p^n$ for some $n \geq 0$. The proof of this result

requires several theorems that are important in themselves and reveal many other properties of groups of p -power order.

Recall that two subgroups H and K of a group G are called **conjugate** in G if $K = gHg^{-1}$ for some $g \in G$. This relation is an equivalence on the set of all subgroups of G , and the analogous equivalence on the elements of G is an important tool in this section. Thus, two elements a and b of a group G are said to be **conjugate** in G if $b = gag^{-1}$ for some $g \in G$. This is an equivalence on G and the equivalence class of $a \in G$ is denoted

$$\text{class } a = \{x \in G \mid x \text{ is conjugate to } a\} = \{gag^{-1} \mid g \in G\},$$

and is called the **conjugacy class** of a .

Hence the conjugacy classes partition a group G . Clearly, $\text{class } 1 = \{1\}$ in any group and, more generally, $\text{class } a = \{a\}$ if and only if $a \in Z(G)$. Also, if a and b are conjugate, then $o(a) = o(b)$ because gag^{-1} is the image of a under an inner automorphism of G . Hence, all elements in a conjugacy class have the same order.

Example 1. Partition D_3 into conjugacy classes.

Solution. Let $D_3 = \{1, a, a^2, b, ba, ba^2\}$, where $o(a) = 3$, $o(b) = 2$, and $aba = b$. We have $\text{class } 1 = \{1\}$. As a and a^2 are the only elements of order 3, we have $\text{class } a \subseteq \{a, a^2\}$. But $a^2 = bab^{-1}$, so $\text{class } a = \{a, a^2\}$. Similarly, $\text{class } b = \{b, ba, ba^2\}$ because $aba^{-1} = ba$ and $a^2ba^{-2} = ba^2$. \square

It can be shown (Exercise 15) that two permutations in S_n are conjugate if and only if they have the same **cycle structure**; that is, when factored into disjoint cycles they have the same number of cycles of each length.

Example 2. The conjugacy classes of S_4 are

$$\begin{aligned} \text{class } \varepsilon &= \{\varepsilon\} \\ \text{class}(1 \ 2) &= \{(1 \ 2), (1 \ 3), (1 \ 4), (2 \ 3), (2 \ 4), (3 \ 4)\} \\ \text{class}(1 \ 2)(3 \ 4) &= \{(1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\} \\ \text{class}(1 \ 2 \ 3) &= \{(1 \ 2 \ 3), (1 \ 3 \ 2), (1 \ 2 \ 4), (1 \ 4 \ 2), \\ &\quad (1 \ 3 \ 4), (1 \ 4 \ 3), (2 \ 3 \ 4), (2 \ 4 \ 3)\} \\ \text{class}(1 \ 2 \ 3 \ 4) &= \{(1 \ 2 \ 3 \ 4), (1 \ 2 \ 4 \ 3), (1 \ 3 \ 2 \ 4), \\ &\quad (1 \ 3 \ 4 \ 2), (1 \ 4 \ 2 \ 3), (1 \ 4 \ 3 \ 2)\} \end{aligned}$$

If K is a normal subgroup of G , then $gKg^{-1} = K$ for all $g \in G$, and so K contains the conjugacy class of each of its elements. Conversely, any subgroup that is a union of conjugacy classes must be normal (Exercise 5). This proves

Theorem 1. If H is a subgroup of a group G , then $H \triangleleft G$ if and only if H is a union of conjugacy classes.

If $D_3 = \{1, a, a^2, b, ba, ba^2\}$, as in Example 1, Theorem 1 shows that any normal subgroup K of D_3 must be a union of the conjugacy classes $\{1\}$, $\{a, a^2\}$, and $\{b, ba, ba^2\}$. Because $1 \in K$ and $|K|$ divides $|D_3| = 6$, the only normal subgroups of D_3 are $\{1\}$, $\{1, a, a^2\}$, and D_3 . Similarly, Example 2 gives Example 3 (Exercise 17).

Example 3. The normal subgroups of S_4 are $\{\varepsilon\}$, A_4 , S_4 , and

$$K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}.$$

The relationship between conjugacy classes and normality is even closer than that shown in Theorem 1. If $X \subseteq G$ is a nonempty subset, write

$$N(X) = N_G(X) = \{g \in G \mid gXg^{-1} = X\}.$$

This is a subgroup of G for every X (Exercise 12), called the **normalizer** of X in G . We write $N(X) = N_G(X)$ if the group G must be emphasized, and we abbreviate $N(\{a\}) = N(a)$ for $a \in G$. Note that $N(a) = \{g \in G \mid ga = ag\}$. For this reason, $N(a)$ is often called the **centralizer** of a in G . The normalizer of a subgroup has the following properties which explain the name.

Lemma 1. *Let H be a subgroup of a group G .*

- (1) $H \triangleleft N(H)$
- (2) *If $H \triangleleft K$, where K is a subgroup of G , then $K \subseteq N(H)$.*

Proof. Let $H \triangleleft K$ and $k \in K$. Then $kHk^{-1} = H$, so $k \in N(H)$. Thus, $K \subseteq N(H)$ proving (2). If we take $K = H$ in (2), we get $H \subseteq N(H)$, whence $H \triangleleft N(H)$. ■

We can summarize Lemma 1 by saying that $N(H)$ is the largest subgroup of G in which H is normal. In particular, $H \triangleleft G$ if and only if $N(H) = G$. At the other extreme, it can happen that $N(H) = H$ (consider $H = \{\varepsilon, (1 \ 2)\}$ in S_3).

Much of the importance of normalizers stems from their connection with conjugation. Recall that $|G : H|$ denotes the index in G of a subgroup $H \subseteq G$.

Theorem 2. *Let G be a finite group.*

- (1) $|\text{class } a| = |G : N(a)|$ for each $a \in G$.
- (2) *The number of conjugates of a subgroup H of G is $|G : N(H)|$.*

Proof. We prove (1); (2) is analogous (Exercise 13). Write $N(a) = N$. The index $|G : N| = |\{gN \mid g \in G\}|$. Since class $a = \{gag^{-1} \mid g \in G\}$, define a mapping

$$\varphi : \text{class } a \rightarrow \{gN \mid g \in G\} \quad \text{by} \quad \varphi(gag^{-1}) = gN.$$

It suffices to prove that φ is a bijection. Now $N = \{x \in G \mid ax = xa\}$, so we have

$$\begin{aligned} gag^{-1} = hah^{-1} &\Leftrightarrow (h^{-1}g)a = a(h^{-1}g) \\ &\Leftrightarrow h^{-1}g \in N \\ &\Leftrightarrow gN = hN \end{aligned}$$

This shows that φ is well defined and one-to-one; as φ is onto, this proves (1). ■

Combining Theorem 2 with the fact that class $a = \{gag^{-1} \mid g \in G\}$ gives

$$a \in Z(G) \Leftrightarrow \text{class } a = \{a\} \Leftrightarrow N(a) = G.$$

In particular, the center $Z(G)$ is the union of all the singleton conjugacy classes. This leads to the following useful theorem.

Theorem 3. The Class Equation. *Let G be a finite group and let class a_1 , class a_2, \dots , class a_n be the nonsingleton conjugacy classes. Then*

$$|G| = |Z(G)| + \sum_{i=1}^n |G : N(a_i)|.$$

Proof. The conjugacy classes partition G , so $|G|$ is the sum of the sizes of these classes. But the number of elements in class a_i is $|G : N(a_i)|$ by Theorem 2, and $|Z(G)|$ is the number of singleton classes. The class equation follows. ■

Example 4. Consider the quaternion group $Q = \{1, -1, i, -i, j, -j, k, -k\}$ as in Example 9 §2.8. The conjugacy classes are $\{1\}$, $\{-1\}$, $\{i, -i\}$, $\{j, -j\}$, and $\{k, -k\}$. We have $N(i) = \{1, -1, i, -i\}$ so that $|Q : N(i)| = 2 = |\text{class } i|$, as in Theorem 2. Because $Z(Q) = \{1, -1\}$, the class equation is apparent.

The class equation is reminiscent of Lagrange's theorem in that it provides arithmetic information about the group. That Lagrange's theorem is useful is beyond doubt; the usefulness of the class equation lies in the fact that each term $|G : N(a)|$ is a divisor of $|G|$ which is not equal to 1 when $a \notin Z(G)$. This fact is particularly useful when $|G|$ is a prime power as we shall see.

However, before doing so we use the class equation to prove an important theorem about general finite groups—due to A. L. Cauchy. If G is a finite group, the order of each element divides $|G|$ by Lagrange's theorem. The converse is false. For example, $|A_4| = 12$ but A_4 has no element of order 6. However, a partial converse does hold.

Theorem 4. Cauchy's Theorem. *If a prime p divides the order of a finite group G , then G has an element of order p .*

Proof. If G is abelian, a (self-contained) proof has already been given (Lemma 1 §7.2). In general, we use induction on $|G|$. The theorem is easily verified if $|G| \leq 3$. If $|G| > 3$, let class a_1, \dots, a_n denote the nonsingleton conjugacy classes so that $|N(a_i)| < |G|$. If p divides $|N(a_i)|$ for any i , the proof is complete by induction. Otherwise, p divides $|G : N(a_i)|$ for each i , and hence p divides $|Z(G)|$ by the class equation. As $Z(G)$ is abelian, Lemma 1 §7.2 completes the proof. ■

As with many important theorems, the method of proof of Cauchy's theorem is at least as important as the result itself. In Section 8.3, we present a sweeping generalization of the class equation, which yields a wealth of information about finite groups.

p-Groups

We use Cauchy's theorem frequently below. One of the most important applications is to characterize groups of prime power order. If p is a prime, a group G is called a **p -group** if the order of every element of G is a power of p .

Lemma 2. *If G is a finite group and p is a prime, then $|G|$ is a power of p if and only if G is a p -group.*

Proof. Assume that $o(g)$ is a power of p for all $g \in G$. If $|G|$ is not a power of p , let q divide $|G|$, where $q \neq p$ is a prime. Then Cauchy's theorem shows that G has an element of order q , contrary to hypothesis. Hence, $|G|$ is a power of p . The converse follows by Lagrange's theorem. ■

Thus, Lemma 2 characterizes the finite p -groups. The next result holds for all p -groups, finite or not, and we leave the routine proof as Exercise 21.

Theorem 5. Let $K \subseteq G$ be groups with $K \triangleleft G$ and let p be a prime. Then G is a p -group if and only if both K and G/K are p -groups.

Although infinite p -groups exist (Exercise 23), we focus on the finite case. Theorem 6 is fundamental, and the proof provides a good illustration of how to use the class equation.

Theorem 6. If p a prime and $G \neq \{1\}$ is a finite p -group, then $Z(G) \neq \{1\}$.

Proof. Let class a_1, \dots, a_n denote the nonsingleton conjugacy classes in G . Because $N(a_i) \neq G$ for each i by Theorem 2, and because $|G : N(a_i)|$ divides $|G|$ for each i , it follows that p divides $|G : N(a_i)|$ for each i . But then p divides $|Z(G)|$ by the class equation; in particular $Z(G) \neq \{1\}$. ■

Theorem 6 is very useful in the study of p -groups where p is a prime. We give two applications; the first characterizes of all groups of order p^2 .

Theorem 7. If G is a group and $|G| = p^2$ where p is a prime, then G is abelian and either $G \cong C_{p^2}$ or $G \cong C_p \times C_p$.

Proof. To prove that G is abelian, we show that $Z(G) = G$. As $Z(G) \neq \{1\}$ by Theorem 6, it suffices to show that $|Z(G)| = p$ is impossible. But, if it holds, then $G/Z(G)$ is cyclic (it has order p), which implies that G is abelian by Theorem 2 §2.9, a contradiction. Hence $|Z(G)| \neq p$, so $Z(G) = G$ and G is abelian. Now assume that G is not cyclic so that every element g satisfies $g^p = 1$. Choose $a \neq 1$ in G and write $H = \langle a \rangle$. Then choose $b \notin H$ and write $K = \langle b \rangle$. Because $|K| = p = |H|$, we have $H \cap K = \{1\}$, so $HK \cong H \times K$ by Theorem 3 §8.1. Hence $|HK| = p^2 = |G|$, so $G = HK \cong H \times K \cong C_p \times C_p$. ■

The extension of Theorem 7 to groups' of order p^3 is false: If $p = 2$, the nonabelian groups D_4 and Q both have order 2^3 . More generally, if p is an odd prime, Exercises 30 and 31 give nonabelian groups G_1 and G_2 of order p^3 such that $g^p = 1$ for all $g \in G_1$, and G_2 contains an element of order p^2 .

The next result shows that, although a finite p -group need not be abelian, it has an abundance of normal subgroups; in fact, it has one of every possible order. The proof again depends on Theorem 6 and provides a tour de force through the methods we have developed for dealing with finite groups.

Theorem 8. Let G be a finite p -group of order p^n . Then there exists a series

$$G = G_0 \supset G_1 \supset \cdots \supset G_n = \{1\}$$

of subgroups of G such that $G_i \triangleleft G$, $|G_i| = p^{n-i}$, and $|G_i/G_{i+1}| = p$ for all i .

Proof. The existence of such a series is obvious if $n = 1$, so we proceed by induction on n . If $|G| = p^{n+1}$, we have $Z(G) \neq \{1\}$ by Theorem 6. By Cauchy's theorem, choose $a \in Z(G)$ such that $o(a) = p$, and write $G_n = \langle a \rangle$. Then $G_n \triangleleft G$ and G/G_n has order p^{n-1} so, by induction, let $(G/G_n) \supset X_1 \supset \cdots \supset X_n = \{G_n\}$ be a series of subgroups of G/G_n such that $X_i \triangleleft G/G_n$ and $|X_i/X_{i+1}| = p$ for each i . The correspondence theorem (Theorem 5 §8.1) ensures that each X_i has the form $X_i = G_i/G_n$, where $G_i \triangleleft G$ and $|G_i/G_n| = p^{(n-1)-i}$. Furthermore, $X_i \supset X_{i+1}$ implies that $G_i \supset G_{i+1}$, and $G_i/G_{i+1} \cong X_i/X_{i+1}$ by the third isomorphism theorem (Theorem 7 §8.1). Hence, $G \supset G_1 \supset \cdots \supset G_n \supset \{1\}$ is the required series for G . ■

Note that Theorem 8 shows that if G is a p -group and $|G| = p^n$, $n \geq 1$, then every subgroup $H \subseteq G$ is contained in a subgroup M with $|M| = p^{n-1}$. Such subgroups must be normal as we shall see in the Corollary to Theorem 1 §8.3, so $G/M \cong C_p$ and M is maximal.

The existence of a series of subgroups such as that in Theorem 8 gives important information about the group. Such series are studied in Chapter 9.

Augustin Louis Cauchy (1789–1857) Cauchy was certainly one of the great mathematicians, and it is said that he and his contemporary Gauss were the last to know all the mathematics of their time. But, unlike Gauss, Cauchy published profusely (surpassed only by Euler and Cayley), and produced 789 papers on topics as diverse as optics, elasticity, differential equations, mechanics, determinants, permutation groups, and probability. He was the effective founder of the theory of functions of a complex variable. In addition he wrote three classic textbooks on analysis in which he firmly established standards of rigor that are now accepted by all analysts and carry down to today's calculus texts. We owe our modern notions of limit and continuity to him. In algebra, Cauchy is remembered as the first to formulate earlier work with permutations in an abstract way and so to create a formal theory of groups of permutations. This work led Cayley (in 1854) to the modern notion of an abstract group.

Cauchy was born in Paris and, after a stellar career in school, enrolled as an engineer in Napoleon's army. He continued his mathematical research and, at the age of 26, became a professor at the École Polytechnique. He soon established himself as the leading mathematician in France. He also enjoyed teaching, and this pedagogical bent probably accounts for the influence his books had.

Exercises 8.2

1. In each case partition G into conjugacy classes and find all the normal subgroups.
 - (a) $G = D_4$
 - (b) $G = Q$
2. Partition D_n into conjugacy classes where n is odd. [Hint: All elements of order 2 are conjugate.]
3. Suppose that $|a| = n$ in a finite group G . If a^m is conjugate to a in G , show that $\gcd(m, n) = 1$. [Hint: If $gag^{-1} = a^m$, show first that $g^2ag^{-2} = a^{m^2}$.]
4. Show that ab and ba are conjugate in any group.
5. If a subgroup H of G is a union of conjugacy classes in G , show that $H \triangleleft G$.
6. If H is a subgroup of prime index in a finite group G , show that either $H \triangleleft G$ or $N(H) = H$.
7. If H and K are conjugate subgroups in G , show that $N(H)$ and $N(K)$ are conjugate.
8. If G is a group, let $K = \langle \text{class } a \rangle$ where $a \in G$. Show that $K \triangleleft G$.
9. If a finite group G has an element with exactly two conjugates, show that G is not simple.
10. If G is a finite group and $H \neq G$ is a subgroup, show that $G \neq \bigcup_{a \in G} aHa^{-1}$. [Hint: Theorem 2.]
11. If H is a subgroup of G of finite index, show that H has only finitely many conjugates in G . [Hint: Exercise 31 §2.6.]
12. Show that $N(X)$ is a subgroup of G for each nonempty subset X of G .
13. Prove (2) of Theorem 2.
14. Let $D_3 = \{1, a, a^2, b, ba, ba^2\}$, where $o(a) = 3$, $o(b) = 2$, and $aba = b$. If $H = \{1, b\}$, show that $N(H) = H$.

15. Use Lemma 3 §2.8 to show that two permutations are conjugate in S_n if and only if they have the same cycle structure.
16. If $\gamma = (1 \ 2 \ 3 \ 4)$ and $\delta = (1 \ 2 \ 3)$ in S_4 , compute $N(\gamma)$ and $N(\delta)$. [Hint: Preceding exercise.]
17. Write $K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$.
 - (a) Show that the only normal subgroups of S_4 are $\{\varepsilon\}$, K , A_4 , and S_4 . [Hint: Exercise 15.]
 - (b) Show that the only normal subgroups of A_4 are $\{\varepsilon\}$, K , and A_4 . [Hint: Exercise 7 §2.8 and Lemma 2 §2.8.]
18. If $n \geq 5$, show that $\{\varepsilon\}$, A_n , and S_n are the only normal subgroups of S_n . [Hint: Theorem 8 §2.8 and Exercise 7 §2.8.]
19. If G is a finite group with exactly two conjugacy classes, show that $|G| = 2$.
20. If G is a group and $a \in G$, define $M(a) = \{g \in G \mid [g, a] \in Z(G)\}$, where we write $[g, a] = gag^{-1}a^{-1}$ for the commutator. Show that $M(a)$ is a subgroup of G and that there is a homomorphism $M(a) \rightarrow Z(G)$ with kernel $N(a)$.
21. Prove Theorem 5.
22. Let G be a finite group. If p is a prime, show that G has a normal subgroup of index p if and only if p divides $|G/G'|$, where G' is the commutator subgroup. [Hint: If p divides $|G/G'|$ apply Theorem 8 to the p -primary component of G/G' .]
23. Let G^ω be the group of sequences $[g_i] = (g_0, g_1, \dots)$ from a group G , with componentwise multiplication $[g_i] \cdot [h_i] = [g_i h_i]$. (See Exercise 37 §2.10). Show that, if $G \neq \{1\}$ is a finite p -group, then G^ω is an infinite p -group.
24. If G is a finite p -group, show that $G' \neq G$. [Hint: Theorem 8.]
25. If $H \triangleleft G$, where G is a finite p -group and $H \neq \{1\}$, show that $H \cap Z(G) \neq \{1\}$. [Hint: Theorem 1.]
26. Let G be a nonabelian group of order p^3 , where p is a prime. Show that
 - (a) $Z(G) = G'$ and this is the unique normal subgroup of G of order p .
 - (b) G has exactly $p^2 + p - 1$ distinct conjugacy classes.
27. Let G be a finite p -group and let $H \triangleleft G$. If $|H| = p^m$ and $|G| = p^n$, strengthen Theorem 8 by showing that a series $G = G_0 \supseteq G_1 \supseteq \dots \supseteq G_n = \{1\}$ exists such that $G_i \triangleleft G$ and $|G_i/G_{i+1}| = p$ for all i , and that $G_{n-m} = H$. [Hint: Exercise 25.]
28. Let G be a group of order p^n and let H_1, \dots, H_m be the distinct subgroups of G of index p . If $N = H_1 \cap \dots \cap H_m$, show that $N \triangleleft G$ and that $x^p = 1$ for every coset x in G/N . [Remark: In fact $H_i \triangleleft G$ for each i (see the Corollary of Theorem 1 §8.3).]
29. If $H \neq G$ is a subgroup of a finite p -group G , show that $H \neq N(H)$. [Hint: If $C = \text{core } H$ (Exercise 26 §2.8), let $Z(G/C) = K/C$, and show that $K \not\subseteq H$ and $K \subseteq N(H)$.]
30. Let $K \triangleleft G$ where G/K is a finite p -group, p a prime. Show that G has a normal subgroup of index p .
31. If p is an odd prime, let $G = \mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$ and define an operation on G by $(x, y, z) \cdot (x_1, y_1, z_1) = (x + x_1, y + y_1, z + z_1 - yx_1)$. Show that G is a nonabelian group of order p^3 in which $a^p = 1$ for all $a \in G$.
32. Let p be a prime and let $X = \{0, p, 2p, \dots, (p-1)p\}$ be the (additive) subgroup of \mathbb{Z}_{p^2} generated by p . Define an operation on the cartesian product $G = X \times \mathbb{Z}_{p^2}$ by $(x, y)(x_1, y_1) = (x + x_1, y + y_1 - yx_1)$. Show that G is a nonabelian group of order p^3 that contains an element of order p^2 .
33. A G is called an **FC -group** if every conjugacy class is finite. Show that
 - (a) If $|G : Z(G)|$ is finite, then G is an FC -group.

- (b) If G is a finitely generated FC -group, show that $|G : Z(G)|$ is finite. [Hint: Exercise 33 §2.6.]
- (c) If $G = \langle X \rangle$, show that G is an FC -group if and only if $|\text{class } x|$ is finite for all $x \in X$.
- (d) Show that every subgroup and image of an FC -group is an FC -group.
- (e) If G is any group, show that $G^* = \{a \in G \mid \text{class } a \text{ is finite}\}$ is an FC -group which is a characteristic subgroup of G .

8.3 GROUP ACTIONS

A mathematician, like a painter or a poet, is a maker of patterns.

—Godfrey Harold Hardy

If G is a finite group of order n , Cayley's theorem asserts that there exists a one-to-one group homomorphism $G \rightarrow S_n$. The proof proceed as follows. Given $a \in G$, we define the multiplication map $\sigma_a : G \rightarrow G$ by $\sigma_a(g) = ag$ for all $g \in G$. We can easily verify that σ_a is a bijection and so belongs to the group S_G of all permutations of the set G . The proof is then completed by observing that the map $G \rightarrow S_G$ given by $a \mapsto \sigma_a$ is a one-to-one homomorphism and so embeds G in the permutation group S_G . (Of course, $S_G \cong S_n$ because $|G| = n$).

The action of the permutation $\sigma_a : G \rightarrow G$ is left multiplication by a . The key observation in this section is that there are sets other than G on which an element of G can act by multiplication. For example, if H is a subgroup of G of index m , let $X = \{gH \mid g \in G\}$ denote the set of all left cosets. Then for $a \in G$ we define

$$\tau_a : X \rightarrow X \quad \text{by} \quad \tau_a(gH) = a(gH) = agH, \quad \text{for all } gH \text{ in } X.$$

One verifies that τ_a is a (well defined) bijection for each $a \in G$ and so $\tau_a \in S_X$. Moreover, $\tau_{ab} = \tau_a \tau_b$ because $\tau_{ab}(gH) = abgH = a(bgH) = \tau_a[\tau_b(gH)]$ for all g . Since this holds for all a and b , the map

$$\varphi : G \rightarrow S_X \quad \text{given by} \quad \varphi(a) = \tau_a, \quad \text{for all } a \in G$$

is a group homomorphism. However, unlike the map in Cayley's theorem, φ may have a nontrivial kernel:

$$\begin{aligned} \ker \varphi &= \{a \in G \mid agH = gH \text{ for all } g \in G\} \\ &= \{a \in G \mid g^{-1}ag \in H \text{ for all } g \in G\} \\ &= \{a \in G \mid a \in gHg^{-1} \text{ for all } g \in G\} \\ &= \bigcap_{g \in G} gHg^{-1}. \end{aligned}$$

This group is important enough to warrant a name.

If H is a subgroup of a group G , the **core** of H in G , denoted $\text{core } H$, is defined to be the intersection of all the conjugates of H in G ; that is,

$$\text{core } H = \{a \in G \mid a \in gHg^{-1} \text{ for all } g \in G\} = \bigcap_{g \in G} gHg^{-1}.$$

Thus, $\text{core } H \triangleleft G$ by the preceding discussion, and $\text{core } H \subseteq H$ because H is a conjugate of itself. Furthermore, $\text{core } H$ is the largest normal subgroup of G that is contained in H . We record this fact for reference, and leave the proof as Exercise 9.

Lemma 1. Let H be a subgroup of a group G .

- (1) $\text{core } H \triangleleft G$ and $\text{core } H \subseteq H$.
- (2) If $K \triangleleft G$ and $K \subseteq H$, then $K \subseteq \text{core } H$.

Our present interest in $\text{core } H$ comes from Theorem 1.

Theorem 1. Extended Cayley Theorem. If H is a subgroup of finite index m in a group G , there is a group homomorphism $\theta : G \rightarrow S_m$ with $\ker \theta = \text{core } H \subseteq H$.

Proof. If $X = \{gH \mid g \in G\}$, let $\varphi : G \rightarrow S_X$ be defined as above. As $|X| = m$, there is an isomorphism $\delta : S_X \rightarrow S_m$, so we obtain a homomorphism $\delta\varphi : G \rightarrow S_m$, and $\ker \delta\varphi = \ker \varphi = \text{core } H$ by the preceding discussion. Then $\theta = \delta\varphi$ does it. ■

This is Cayley's theorem when $H = \{1\}$. Example 1 illustrates how to use it.

Example 1. If $|G| = 36$ and G has a subgroup H of order 9,⁸⁹ then G is not simple. Indeed, $|G : H| = 4$ so, by Theorem 1, there is a homomorphism $\theta : G \rightarrow S_4$, with $\ker \theta \subseteq H$. If $\ker \theta = \{1\}$, then $G \cong \theta(G)$, a contradiction as $|G| = 36$ and $|\theta(G)| \leq |S_4| = 24$. So $\ker \theta \neq \{1\}$ is normal in G .

In Section 2.8, we showed that any subgroup of index 2 is normal. The next result gives a generalization that is especially useful for finite p -groups.

Corollary. Let p be the smallest prime dividing the order of a finite group G . Then any subgroup of G of index p is normal in G .

Proof. Let $|G| = p^k q^m r^n \dots$, where $p < q < r \dots$ are primes. If $|G : H| = p$, then $|H| = p^{k-1} q^m r^n \dots$. By Theorem 1 let $\theta : G \rightarrow S_p$ be a homomorphism with $\ker \theta \subseteq H$ and write $K = \ker \theta$. If $|K| = p^{k-1-k_0} q^{m-m_0} r^{n-n_0} \dots$, then we have $|G/K| = p^{1+k_0} q^{m_0} r^{n_0} \dots$ divides $|S_p| = p!$ and so $p^{k_0} q^{m_0} r^{n_0} \dots$ divides $(p-1)!$. But this implies that $k_0 = m_0 = n_0 = \dots = 0$ because every divisor of $(p-1)!$ is less than p . Hence, $H = K \triangleleft G$. ■

We give more applications of the extended Cayley theorem later; our present aim is to generalize it. The key to the theorem is the existence of the homomorphism $\varphi : G \rightarrow S_X$, where G is a group and X is some set. Because the image $\varphi(G)$ of G is a subgroup of S_X , the natural place to begin is to consider this situation.

Hence suppose that X is a nonempty set and that G is a subgroup of the group S_X of all permutations of X . For $x \in X$ and $\sigma \in G$, the element $\sigma(x)$ of X is specified, which amounts to a mapping $G \times X \rightarrow X$ where $(\sigma, x) \mapsto \sigma(x)$. We can now describe an apparently more general situation.

Group Actions

Let G be a group and let X be a nonempty set. A mapping $G \times X \rightarrow X$, denoted $(a, x) \mapsto a \cdot x$, is called an **action** of G if it satisfies the following conditions.

- A1 $1 \cdot x = x$ for all $x \in X$.
- A2 $a \cdot (b \cdot x) = (ab) \cdot x$ for all $x \in X$ and for all $a, b \in G$.

In this case, G is said to **act** on X and X is called a **G -set**.⁹⁰

⁸⁹Such a subgroup H must in fact exist (see Theorem 1 §8.4).

⁹⁰An action on the right may be defined by $(a, x) \mapsto x * a \in X$. This is nothing new because $a \cdot x = x * a^{-1}$ is then an action in the present sense.

Hence an action of G on X is nothing more than a *multiplication* of any element x of X by any element a of G to yield a (uniquely determined) element $a \cdot x$ of X that satisfies axioms A1 and A2. There are many examples of such actions, and Example 2 recaptures the above discussion.

Example 2. If X is any nonempty set and $G \subseteq S_X$ is any group of permutations of X , define $\sigma \cdot x = \sigma(x)$ for all $x \in X$ and $\sigma \in G$. Then axioms A1 and A2 are clearly satisfied; in fact, A2 is the definition of composition of mappings.

Example 3. Let H be a subgroup of a group G . Consider G as a set for the moment and let H act on G by $h \cdot x = hx$ for all $x \in G$, $h \in H$. This is clearly an action; and H is said to **act on G by left multiplication**.

Example 4. If G is a group, let G act on itself by $a \cdot x = axa^{-1}$ for all $x \in G$ and $a \in G$. Then axiom A1 is clear and A2 holds because

$$a \cdot (b \cdot x) = a(bxb^{-1})a^{-1} = (ab)x(ab)^{-1} = (ab) \cdot x$$

for $x \in G$ and $a, b \in G$. In this case, G is said to **act on itself by conjugation**.

Example 5. Let H be a subgroup of a group G and let $X = \{gH \mid g \in G\}$ denote the set of left cosets of H in G . If $a \in G$ and $gH \in X$ then $a \cdot (gH) = agH$ is well defined (verify) and so is an action of G on X . As we have shown, this action plays an essential role in the derivation of the extended Cayley theorem.

Example 6. If X is any set and G is any group, we define the **trivial action** by $a \cdot x = x$ for all $x \in X$ and $a \in G$. Clearly, the axioms are satisfied.

Examples 2–6 show that group actions are commonly occurring phenomena, and other examples below underline this conclusion. Lemma 2 isolates two useful properties of group actions that we use repeatedly.

Lemma 2. Let X be a G -set, where G is a group, and let $x, y \in X$ and $a, b \in G$.

- (1) If $a \cdot x = a \cdot y$, then $x = y$.
- (2) $a \cdot x = b \cdot y$ if and only if $(b^{-1}a) \cdot x = y$.

Proof. Clearly, (1) follows from (2) and axiom A1. If $a \cdot x = b \cdot y$, then

$$(b^{-1}a) \cdot x = b^{-1} \cdot (a \cdot x) = b^{-1} \cdot (b \cdot y) = (b^{-1}b) \cdot y = 1 \cdot y = y,$$

which proves half of (2). The other implication is proved similarly. ■

We can now give a natural generalization of the extended Cayley theorem. If G is a group, X is a G -set and $a \in G$, define

$$\sigma_a : X \rightarrow X \quad \text{by} \quad \sigma_a(x) = a \cdot x \text{ for all } x \in X. \tag{*}$$

Then Lemma 2 shows that σ_a is a bijection and so is a member of the group S_X of all permutations of X . Moreover, if $a, b \in G$, then axiom A2 gives

$$\sigma_{ab}(x) = (ab) \cdot x = a \cdot (b \cdot x) = \sigma_a[\sigma_b(x)] = (\sigma_a \sigma_b)(x)$$

for all $x \in S$. Hence, $\sigma_{ab} = \sigma_a \sigma_b$ so the map $\theta : G \rightarrow S_X$, defined by $\theta(a) = \sigma_a$ for all $a \in G$, is a group homomorphism. This gives parts (1) and (2) of

Theorem 2. Let G be a group, let X be a G -set, and let σ_a be defined as in (*), where $a \in G$. Then

- (1) $\sigma_a \in S_X$ for all $a \in G$.
- (2) $\theta : G \rightarrow S_X$ given by $\theta(a) = \sigma_a$ is a group homomorphism.
- (3) $\ker \theta = \{a \in G \mid a \cdot x = x \text{ for all } x \in X\}$.

Proof. It remains to prove (3). But $a \in \ker \theta$ means $\sigma_a = 1_{S_X}$; that is, $\sigma_a(x) = x$ for all $x \in X$. This condition means that $a \cdot x = x$ for all x , as required. \blacksquare

If H is a subgroup of G , the extended Cayley theorem is clearly the special case of Theorem 2 where $X = \{gH \mid g \in G\}$ and the action of G on X is given by $a \cdot (gH) = agH$ for all $gH \in X$ and $a \in G$. In this case $\ker \theta = \text{core } H$. In general, if X is a G -set then $\ker \theta \triangleleft G$ is given by

$$\ker \theta = \{a \in G \mid a \cdot x = x \text{ for all } x \in X\}.$$

The normal subgroup $\ker \theta$ is called the **fixer** of the action. The reason for the name is that $\ker \theta$ consists of the elements of G that fix every $x \in X$. Here an element $x \in X$ is said to be **fixed** by $a \in G$ if $a \cdot x = x$.

Example 7. Let G be a group and let G act on itself by conjugation: $a \cdot x = axa^{-1}$ for all $x \in G$ and $a \in G$. Here the bijection $\sigma_a : G \rightarrow G$ in Theorem 2 is the inner automorphism of G induced by a . Hence $\theta(G) = \text{inn } G$, where $\theta : G \rightarrow S_G$ is the homomorphism in Theorem 2, and the fixer is $\ker \theta = \{a \in G \mid a \cdot x = x \text{ for all } x \in G\} = Z(G)$ in this case. Thus, Theorem 2 gives $G/Z(G) \cong \text{inn } G$, a result derived earlier (Theorem 5 §2.10).

Orbit Decomposition Theorem

So far, the theory of group actions has been motivated by the urge to generalize the extended Cayley theorem. However, the theory yields an additional bonus: It provides a fundamental, natural generalization of the class equation.

Recall that elements x and y in a group G are called conjugate in G if $y = axa^{-1}$ for some $a \in G$. If we regard G as acting on itself by conjugation, this condition is $y = a \cdot x$ for some $a \in G$. This suggests a generalization: If X is a G -set, we define a relation \equiv on X as follows. If $x, y \in X$, we write

$$x \equiv y \pmod{G}, \quad \text{if } y = a \cdot x \text{ for some } a \in G.$$

One easily verifies that \equiv is an equivalence on X (Exercise 17). Moreover, we may describe the equivalence class $[x]$ of an element x of X in terms of the action: $[x] = \{y \in X \mid y \equiv x\} = \{a \cdot x \mid a \in G\}$. This equivalence class is called the **orbit** of x under G and is denoted

$$G \cdot x = \{a \cdot x \mid a \in G\}.$$

Hence, if G acts on itself by conjugation, the orbits are just the conjugacy classes. Cosets also occur as orbits.

Example 8. Let H be a subgroup of a group G and let H act on G by left multiplication: $h \cdot x = hx$ for all $x \in G$, $h \in H$. Then the orbit of $x \in G$ is the right coset $Hx = H \cdot x$.

A key step in the derivation of the class equation is the observation that the number of elements in the conjugacy class of $x \in G$ is just the index in G of the normalizer $N(x)$ of x in G . Surprisingly, if X is *any* G -set, the size of each orbit is the index of a certain subgroup of G .

Lemma 3. *If X is a G -set and $x \in X$, write $S(x) = \{a \in G \mid a \cdot x = x\}$. Then*

- (1) *$S(x)$ is a subgroup of G for each $x \in X$.*
- (2) *$|G \cdot x| = |G : S(x)|$ for each $x \in X$.*

Proof. The proof of (1) is left as Exercise 23. Given $x \in X$, write $S(x) = S$ and define a function $\varphi : G \cdot x \rightarrow \{gS \mid g \in G\}$ by $\varphi(g \cdot x) = gS$. Then

$$g \cdot x = h \cdot x \Leftrightarrow (h^{-1}g) \cdot x = x \Leftrightarrow h^{-1}g \in S \Leftrightarrow gS = hS,$$

so φ is well defined and one-to-one. Since φ is clearly onto, this proves (2) because $|G : S| = |\{gS \mid g \in G\}|$. ■

If X is a G -set and $x \in X$, the subgroup

$$S(x) = \{a \in G \mid a \cdot x = x\}$$

is called the **stabilizer**⁹¹ of x . If G acts on itself by conjugation the stabilizer of x in G is $S(x) = \{a \in G \mid axa^{-1} = x\} = N(x)$ is the normalizer of x , and the orbit of x is $G \cdot x = \{gxg^{-1} \mid g \in G\} = \text{class } x$. Hence, Lemma 3 gives $|\text{class } x| = |G : N(x)|$ in this case, a result proved earlier (Theorem 2 §8.2).

If X is a G -set and $x \in X$, the orbit of x is $G \cdot x = \{a \cdot x \mid a \in G\}$. Combining this with Lemma 3 gives equivalent conditions that the orbit is a singleton

$$G \cdot x = \{x\} \Leftrightarrow a \cdot x = x, \text{ for all } a \in G \Leftrightarrow S(x) = G. \quad (**)$$

The set of all such elements x is denoted

$$X_f = \{x \in X \mid a \cdot x = x \text{ for all } a \in G\}$$

and is called the **fixed subset** of X under the action of G . With this we can give the promised generalization of the class equation.

Theorem 3. Orbit Decomposition Theorem. *Let a group G act on a finite set $X \neq \emptyset$ and let $G \cdot x_1, G \cdot x_2, \dots, G \cdot x_n$ denote the nonsingleton orbits. Then*

$$|X| = |X_f| + \sum_{i=1}^n |G : S(x_i)|.$$

Proof. The fixed subset X_f is the union of the singleton orbits by (**). Because the orbits partition X , $|X| = |X_f| + \sum_{i=1}^n |G \cdot x_i|$. Now apply Lemma 3. ■

Theorem 3 becomes the class equation if $X = G$ and G acts on itself by conjugation because the fixed subset is $G_f = \{x \in G \mid axa^{-1} = x \text{ for all } a \in G\} = Z(G)$.

In the terminology of Theorem 3 the index $|G : S(x_i)| = |G \cdot x_i|$ is finite because X is a finite set and, if the group G is itself finite, it is a divisor of $|G|$. This property is particularly important when G is a finite p -group, where p is a prime, because then p divides $|G : S(x_i)|$ for each i . Hence, Theorem 3 shows that p divides $|X| - |X_f|$ in this case, which is important enough to record as Theorem 4.

⁹¹Another name for $S(x)$ is the **isotropy group** of x .

Theorem 4. Let p be a prime and let G be a finite p -group. If X is a finite G -set, then p divides $|X| - |X_f|$.

We use this result repeatedly in the next Section. For now we illustrate how to use it by proving an important property of finite p -groups that will recur in Chapter 9.

Theorem 5. Let G be a finite p -group, where p is a prime. If $H \neq G$ is a subgroup, then $N(H) \neq H$.

Proof. Let $X = \{xH \mid x \in G\}$ denote the set of left cosets of H in G , and let H act on X by left multiplication: $h \cdot (xH) = hxH$ for all $x \in G$ and $h \in H$. Then p divides $|X| = |G : H|$ and so $|X_f| \neq 1$ by Theorem 4. Now

$$\begin{aligned} X_f &= \{xH \mid hxH = xH \text{ for all } h \in H\} \\ &= \{xH \mid x^{-1}Hx \subseteq H\} \\ &= \{xH \mid x \in N(H)\}. \end{aligned}$$

Hence, $N(H) \neq H$ since otherwise $X_f = \{H\}$, contrary to $|X_f| \neq 1$. ■

We conclude this section by sketching J.H. McKay's beautiful proof⁹² of Cauchy's theorem, which applies Theorem 4 and avoids proving the abelian case separately. If G is a group and p is a prime divisor of $|G|$, we must find an element of order p in G . McKay's idea is to consider the set of p -tuples with product 1:

$$X = \{(a_1, \dots, a_p) \mid a_i \in G, a_1 a_2 \cdots a_p = 1\}.$$

What is needed is a p -tuple (a, a, \dots, a) in X with $a \neq 1$. To this end, let the (additive) group \mathbb{Z}_p act on X by “cycling” the p -tuples:

$$\bar{k} \cdot (a_1, \dots, a_p) = (a_{1+k}, \dots, a_p, a_1, \dots, a_k), \quad \text{for all } \bar{k} \in \mathbb{Z}_p.$$

We leave to the reader the task of verifying that this is an action (well defined) and that the fixed subset is $X_f = \{(a, a, \dots, a) \mid a^p = 1\}$. Hence, Cauchy's theorem follows if $|X_f| \neq 1$, and this in turn holds (by Theorem 4) if p divides $|X|$. But this latter condition follows because p divides $|G|$ and $|X| = |G|^{p-1}$ (indeed, in choosing (a_1, \dots, a_p) in X , the elements a_1, \dots, a_{p-1} can be selected arbitrarily from G). This completes a most elegant proof.

Exercises 8.3

1. (a) If $|G| = 20$, show that G has a normal subgroup of order 5.
(b) If $|G| = 28$, show that G has a normal subgroup of order 7.
2. If $|G| = 24$ and G has a subgroup of order 8, show that G is not simple.
3. If p and q are primes, show that no group of order pq is simple.
4. Show that every group of order 15 is cyclic. [Hint: If $o(a)=5$ and $o(b)=3$, show that $bab^{-1} = a^k$ for some k . Deduce that $b^nab^{-n} = a^{kn}$ for each n and hence $k = 1$.]
5. If $|G| = pm$, where p is a prime and $p > m$, show that any subgroup of order p is normal in G . (Such subgroups exist by Cauchy's theorem.)
6. (a) If $n \geq 5$ and $p \neq n$ is a prime, show that A_n has no subgroup of index p .
(b) If p is a prime, show that A_p has a subgroup of index p .

⁹²American Mathematical Monthly, Vol. 66, 1956, p. 119.

7. If H and K are subgroups of G , show that $\text{core}(H \cap K) = \text{core } H \cap \text{core } K$.
8. If G is the group of all 2×2 invertible matrices over \mathbb{R} , find $\text{core } H$, where H is the group of diagonal matrices $\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$ in G .
9. Prove Lemma 1.
10. If H is a subgroup of G , define $H_0 = \bigcap\{\sigma(H) \mid \sigma \in \text{aut } G\}$.
 - (a) Show that H_0 is characteristic in G and $H_0 \subseteq H$.
 - (b) If K is characteristic in G and $K \subseteq H$, show that $K \subseteq H_0$.
11. Show that the following are equivalent for a group G .
 - (1) G has a nontrivial finite G -set.
 - (2) G has a proper normal subgroup of finite index.
 - (3) G has a proper subgroup of finite index.
12. Given $m > 1$, show that a finitely generated group G has at most a finite number of subgroups of index m . [Hint: If $\mathcal{C} = \{K \mid K = \text{core } H \text{ where } |G : H| = m\}$, show that \mathcal{C} is a finite set and that, given $K \in \mathcal{C}$, there are at most a finite number of subgroups H with $K \subseteq H$.]
13. Let $G = (\mathbb{R}, +)$ and define $a \cdot z = e^{ia}z$ for all $z \in \mathbb{C}$ and $a \in G$. Show that \mathbb{C} is a G -set, describe the action geometrically, and find all orbits and stabilizers.
14. Let $X = \mathbb{R}[x_1, x_2, \dots, x_n]$, the polynomial ring in the indeterminates x_1, \dots, x_n . Given $\sigma \in S_n$ and $f = f(x_1, \dots, x_n) \in X$, define $\sigma \cdot f = f(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n})$. Show that this is an action and describe the fixer. If $n = 3$, give three polynomials in the fixer and compute $S_3 \cdot g$ and $S(g)$, where $g(x_1, x_2, x_3) = x_1 + x_2$.
15. Write $X_n = \{1, 2, \dots, n\}$. If $\sigma \in S_n$, write $G = \langle \sigma \rangle$ and let the elements of G act on X_n as mappings. Describe the relationship between the orbits of G in X_n and the factorization of σ into disjoint cycles.
16. Let $\theta : G \rightarrow S_X$ be a group homomorphism, where X is a nonempty set. Show that θ arises, as in Theorem 2, from some action of G on X .
17. Let X be a G -set. Show that
 - (a) Equivalence modulo G (defined prior to Example 8) is an equivalence on X .
 - (b) Every equivalence on X arises, as in (a), from some group action on X .
18. Let X be a G -set. If F is the fixer, show that X is a G/F -set in a natural way and that the fixer is trivial (such actions are called **faithful**).
19. Show that a group G acts on its set of subgroups by conjugation and that $Z(G) \subseteq F$, where F is the fixer. Give an example where $Z(G) \neq F$.
20. Is every normal subgroup of a group G the fixer of some action of G ? Give reasons.
21. If H is a subgroup of G , find a G -set X and an element $x \in X$ such that $H = S(x)$.
22. If $H \triangleleft G$, define the **centralizer** of H in G as $C(H) = \{a \in G \mid ah = ha \text{ for all } h \in H\}$. Use Theorem 2 to show that $C(H) \triangleleft G$ and that $G/[C(H)]$ is isomorphic to a group of automorphisms of H .
23. Let X be a G -set and let x and y denote elements of X .
 - (a) Show that $S(x)$ is a subgroup of G .
 - (b) If $x \in X$ and $b \in G$, show that $S(b \cdot x) = bS(x)b^{-1}$.
 - (c) If $S(x)$ and $S(y)$ are conjugate subgroups, show that $|G \cdot x| = |G \cdot y|$.
24. Let X be a G -set with just one orbit (called a **transitive action**).
 - (a) If $K \triangleleft G$, show that $K \subseteq S(x)$ for some $x \in X$ if and only if K is contained in the fixer. [Hint: Exercise 23.]
 - (b) If $|X| \geq 2$, show that $g \in G$ exists such that $g \cdot x \neq x$ for all $x \in X$. [Hint: Exercise 10 §8.2.]

25. Let X be a G -set, let H be a subgroup of G , and let $x \in X$. Show that:
- H acts on the orbit $G \cdot x$ by $h \cdot (a \cdot x) = (ha) \cdot x$ for all $h \in H$ and $a \cdot x \in G \cdot x$.
 - If $H \triangleleft G$, the orbits of H in $G \cdot x$ all have the same cardinality.
26. Let G be a finite p -group. If $\{1\} \neq H \triangleleft G$, show that $H \cap Z(G) \neq \{1\}$. [Hint: Let G act on H by conjugation.]
27. If G is a finite p -group, show that the number of nonnormal subgroups of G is a multiple of p . [Hint: Let G act on its subgroups by conjugation.]
28. If G is a finite p -group, show that the number of subgroups of order p^k is congruent (modulo p) to the number of normal subgroups of order p^k .
29. Let H_1, \dots, H_m be all the subgroups of index p in a finite p -group G . Show that $K = \bigcap_{i=1}^m H_i$ is normal in G , that G/K is abelian, and that $o(x) = p$ for all nonunity elements $x \in G/K$. [Hint: Theorem 5 and Exercise 6 §8.2.]
30. Let H and K be subgroups of a group G . Show that K has $|H : H \cap N(K)|$ distinct conjugates of the form hKh^{-1} , where $h \in H$. Here $N(K) = N_G(K)$ is the normalizer.
31. If H and K are finite subgroups of some group, prove that $|HK| \cdot |H \cap K| = |H| \cdot |K|$ by letting $H \curvearrowright K$ act on $H \times K$ by $a \cdot (h, k) = (ha^{-1}, ak)$. [Hint: Show that each orbit has the same number of elements.]
32. Let H and K be subgroups of a group G and let $H \times K$ act on G by $(h, k) \cdot x = h x k^{-1}$ for all $x \in G$ and $(h, k) \in H \times K$. Show
 - This is an action and the orbit of $x \in G$ is HxK (called a **double coset**).
 - If $x \in G$, then $|S(x)| = |H \cap xKx^{-1}| = |x^{-1}Hx \cap K|$.
 - Frobenius theorem:** If $Hx_1K, Hx_2K, \dots, Hx_nK$ are the distinct double cosets, then $|G| = \sum_{i=1}^n \frac{|H||K|}{|x_i^{-1}Hx_i \cap K|}$.
33. If X and Y are G -sets, a map $\varphi : X \rightarrow Y$ is called a **G -morphism** if $\varphi(a \cdot x) = a \cdot \varphi(x)$ holds for all $x \in X$ and $a \in G$. If, in addition, φ is a bijection, it is called a **G -isomorphism**, and X and Y are called **isomorphic G -sets**. Call a G -set transitive if there is just one orbit. If H is a subgroup of G , let G/H denote the G -set of left H -cosets using left multiplication. (H need not be normal.)
 - Show that G/H is transitive for any subgroup H of G .
 - Show that every transitive G -set is G -isomorphic to G/H for some subgroup H .
34. Let $\varphi : X \rightarrow Y$ be an onto G -morphism, where X and Y are G -sets (Exercise 33). Define a relation \sim on X by $x \sim x_1$ if $\varphi(x) = \varphi(x_1)$. This relation is an equivalence (called the *kernel equivalence of φ*), and we denote the equivalence class of $x \in X$ by $[x] = \{t \in X \mid \varphi(t) = \varphi(x)\}$. Finally, let $X/\varphi = \{[x] \mid x \in X\}$ denote the set of equivalence classes.
 - Show that X/φ is a G -set via $a \cdot [x] = [a \cdot x]$ for all $[x]$ in X/φ and all $a \in G$.
 - Find a G -isomorphism $X/\varphi \rightarrow Y$ (the G -set isomorphism theorem).

8.4 THE SYLOW THEOREMS

Lagrange's theorem asserts that the order of each subgroup of a finite group G is a divisor of $|G|$. The converse is false: A_4 has no subgroup of order 6 even though $|A_4| = 12$. However, if p^k divides the order of G where p is a prime, then G *does* have a subgroup of order p^k . This remarkable theorem was first proven in 1872 (for permutation groups) by the Norwegian mathematician Ludwig Sylow and has been ranked with Lagrange's theorem as being among the most important results about

finite groups. The version presented here for abstract groups was proven in 1887 by Georg Frobenius, and this proof uses only Cauchy's theorem and the class equation. We give another more modern direct proof using the theory of group actions at the end of this section.

Theorem 1. *Let G be a finite group. If p is a prime and p^k divides $|G|$ for some $k \geq 0$, then G has a subgroup of order p^k .*

Proof. It is clear if $k = 0$, so assume $k \geq 1$. Proceed by induction on $|G|$. The theorem is clear if $|G| = 1, 2$, or 3 . In general, $|G| = |Z(G)| + \sum_{i=1}^n |G : N(a_i)|$ by the class equation, where class a_1, \dots, a_n are the nonsingleton conjugacy classes. If p^k divides $|N(a_i)|$ for some i then $N(a_i)$ has a subgroup of order p^k by induction because $|N(a_i)| < |G|$.

So assume that p^k does not divide $|N(a_i)|$ for every $i \geq 1$. Because p^k divides $|G| = |N(a_i)||G : N(a_i)|$ for all i , it follows that p divides $|G : N(a_i)|$ for every i . Hence, p divides $|Z(G)|$ by the class equation so, by Cauchy's theorem, choose $a \in Z(G)$ with $o(a) = p$. If we write $K = \langle a \rangle$ then $K \triangleleft G$ and $|G/K| = \frac{1}{p}|G| < |G|$. Moreover, p^{k-1} divides $|G/K|$ because p^k divides $|G|$. Hence, again by induction, G/K has a subgroup H/K with $|H/K| = p^{k-1}$. As $|H| = p^k$, we are done. ■

Sylow originally proved Theorem 1 in the special case where p^k is the highest power of the prime p that divides the order of the group.

Corollary. Sylow's First Theorem. *If G is a group of order $p^n m$, where p is a prime and p does not divide m , then G has a subgroup of order p^n .*

If G is a group of order $p^n m$, where p is a prime and p does not divide m , any subgroup of order p^n is called a **Sylow p -subgroup** of G .

Example 1. Write $D_3 = \{1, a, a^2, b, ba, ba^2\}$, where $o(a) = 3$, $o(b) = 2$, and $aba = b$. Then $H = \{1, a, a^2\}$ is the unique Sylow 3-subgroup, but $\{1, b\}$, $\{1, ba\}$, and $\{1, ba^2\}$ are three Sylow 2-subgroups. Hence, the Sylow 2-subgroups may not be normal or unique. We show later that a Sylow p -subgroup is normal if and only if it is unique.

Example 2. If G is a finite abelian group and p is a prime divisor of $|G|$, let $G(p) = \{a \in G \mid o(a) = p^k \text{ for some } k \geq 0\}$. This set is a subgroup (because G is abelian) and so is a p -subgroup of G that contains every p -subgroup. It is thus the unique Sylow p -subgroup of G , called the **p -primary component** of G .

Corollary 2 of Theorem 3 §7.2 shows that every finite abelian group is isomorphic to the direct product of its primary components and thus of its distinct Sylow p -subgroups. We characterize when this happens in a nonabelian group in Section 9.3.

A Sylow p -subgroup P of a finite group G is a p -subgroup of G of maximum possible order (by Theorem 1). Note that each conjugate aPa^{-1} of P is also a Sylow p -subgroup because $|aPa^{-1}| = |P|$. The converse is also true: Every Sylow p -subgroup is conjugate to P . In fact, *every* p -subgroup of G is contained in a conjugate of P (and so is contained in a Sylow p -subgroup of G)

Theorem 2. *Let P be a Sylow p -subgroup of a finite group G . If H is any p -subgroup of G , then $H \subseteq aPa^{-1}$ for some $a \in G$.*

Proof. Let $X = \{aP \mid a \in G\}$ be the set of left cosets of P in G and let H act on X by left multiplication: $h \cdot aP = haP$ for all $h \in H$. Write $|G| = p^n m$, where p does not divide m . Then $|X| = |G : P| = m$, so p does not divide $|X|$. Hence (because H is a p -group) Theorem 4 §8.3 shows that p does not divide $|X_f|$, where X_f is the fixed subset. In particular, X_f is not empty, so let $aP \in X_f$, $a \in G$. Then $haP = h \cdot aP = aP$ for all $h \in H$, whence $a^{-1}ha \in P$ for all $h \in H$. Thus $H \subseteq aPa^{-1}$, as required. \blacksquare

Taking H to be any Sylow p -subgroup of G , we obtain

Corollary 1. Sylow's Second Theorem. *If G is a finite group, any two Sylow p -subgroups of G are conjugate in G .*

Because a subgroup of G is normal in G if and only if it equals all its conjugates in G , we get Corollary 2.

Corollary 2. *A Sylow p -subgroup of a finite group G is normal in G if and only if it is unique.*

Example 3. Given D_3 , as in Example 1, the Sylow 2-subgroups $\{1, b\}$, $\{1, ba\}$, and $\{1, ba^2\}$ must be conjugate. In fact, $a\{1, b\}a^{-1} = \{1, ba\}$ and $a^2\{1, b\}a^{-2} = \{1, ba^2\}$.

The next result identifies a fundamental technique due to Giovanni Frattini. We return to this in Section 9.3.

Corollary 3. Frattini Argument. *Let $H \triangleleft G$ and let P be a Sylow p -subgroup of H . Then we have $G = HN_G(P)$, where $N_G(P)$ is the normalizer of P in G .*

Proof. Suppose $g \in G$. Then $gPg^{-1} \subseteq H$ because $H \triangleleft G$, so gPg^{-1} is also a Sylow p -subgroup of H . By Corollary 1, $h(gPg^{-1})h^{-1} = P$ for some $h \in H$, so $hg \in N_G(P)$. The result follows. \blacksquare

If $K \triangleleft G$, Example 4, employs Theorem 2 to provide some useful information about the Sylow subgroups of K and G/K .

Example 4. Let $K \triangleleft G$, where G is a finite group, and let p be a prime.

- (1) A subgroup of K is a Sylow p -subgroup of K if and only if it has the form $P \cap K$, where P is a Sylow p -subgroup of G .
- (2) A subgroup of G/K is a Sylow p -subgroup of G/K if and only if it has the form $(PK)/K$, where P is a Sylow p -subgroup of G .

Solution.

- (1) We begin with one of the implications in (1).

Claim: If P is a Sylow p -subgroup of G then $P \cap K$ is a Sylow p -subgroup of K .

Proof. $P \cap K$ is a p -subgroup of K , so let $P \cap K \subseteq X$ where X is a Sylow p -subgroup of K . But X is a p -subgroup of G so $X \subseteq aPa^{-1}$ for some $a \in G$ by Theorem 2. It follows that

$$a^{-1}(P \cap K)a \subseteq a^{-1}Xa \subseteq P \cap (a^{-1}Ka) = P \cap K.$$

Hence, these sets are equal, and in particular $P \cap K = X$. This proves the Claim.

Now if H is a Sylow p -subgroup of K , then $H \subseteq P$ for some Sylow p -subgroup P of G . Hence, $H \subseteq P \cap K$, so $H = P \cap K$ by the Claim. This proves (1).

(2) Let $|G| = p^n m$ where p does not divide m . By Lagrange's theorem, let $|K| = p^k r$ where $k \leq n$ and $r \mid m$. Thus $\left|\frac{G}{K}\right| = p^{n-k} \left(\frac{m}{r}\right)$. Let $\frac{H}{K}$ be some Sylow p -subgroup of $\frac{G}{K}$ so $\left|\frac{H}{K}\right| = p^{n-k}$. Then $|H| = p^n r$. So if P is a Sylow p -subgroup of H , then P is a Sylow p -subgroup of G , and it remains to show $PK = H$. But $\left|\frac{PK}{K}\right| = \frac{|P|}{|P \cap K|}$ by the second isomorphism theorem (Theorem 1 §8.1), and $|P \cap K| = p^k$ by (1). Hence, $\left|\frac{PK}{K}\right| = p^{n-k}$ so $|PK| = p^n r = |H|$. Since $PK \subseteq H$, this gives $PK = H$, as required. The proof that each of the groups $\frac{PK}{K}$ is a Sylow p -subgroup of $\frac{G}{K}$ is left to the reader. \square

The third Sylow theorem is concerned with the number of Sylow p -subgroups of a finite group G where p is a prime. We will denote this by n_p or $n_p(G)$ if the group must be identified. Although determining n_p from the order of the group is not possible in general, we can deduce a good deal of numerical information.

Theorem 3. Sylow's Third Theorem. *Let G be a group of order $p^n m$, where p is a prime, $n \geq 1$, and p does not divide m . If n_p denotes the number of distinct Sylow p -subgroups of G , then:*

- (1) $n_p \equiv 1 \pmod{p}$.
- (2) n_p divides m .
- (3) $n_p = |G : N(P)|$, where P is any Sylow p -subgroup of G .

Proof. By Sylow's second theorem, (3) follows by Theorem 2 §8.2, so n_p divides $|G| = p^n m$. Hence, (2) follows from (1) because (1) implies that $\gcd(p, n_p) = 1$.

To prove (1), let X denote the set of all Sylow p -subgroups of G so that $|X| = n_p$. Fix P in X and let P act on X by conjugation. If X_f is the fixed subset, then $n_p = |X| \equiv |X_f|$ modulo p by Theorem 4 §8.3, so it suffices to show that $X_f = \{P\}$. We have $X_f = \{Q \in X \mid aQa^{-1} = Q \text{ for all } a \in P\}$, so $P \in X_f$ is clear. If $Q \in X_f$, then $P \subseteq N(Q)$, so both P and Q are Sylow p -subgroups of $N(Q)$ (they are p -subgroups of maximal order). But $Q \triangleleft N(Q)$, so $Q = P$ follows by Corollary 2 of Theorem 2. Hence $X_f = \{P\}$, as required. \blacksquare

Examples 5–9 illustrate the power of the Sylow theorems and how to apply them to particular groups.

Example 5. If p and q are primes, show that no group G of order pq is simple.

Solution. If $p = q$, then G is abelian by Theorem 7 §8.2, so G is not simple because $|G| = p^2$ is not a prime. So assume that $p > q$. Then $n_p \equiv 1 \pmod{p}$ and $n_p \mid q$ by Theorem 3, so $n_p = 1$ because $q < p$. Thus, there is just one Sylow p -subgroup, and it is normal by Corollary 2 of Theorem 2. \square

Example 6. Show that every group of order 175 is abelian.

Solution. Observe $|G| = 175 = 5^2 \cdot 7$. Then $n_5 \mid 7$ and $n_5 \equiv 1 \pmod{5}$ by Theorem 3, from which $n_5 = 1$. Hence, there is just one Sylow 5-subgroup P of G and so $P \triangleleft G$. Similarly $n_7 \mid 5^2$ so $n_7 = 1, 5$, or 25, and $n_7 \equiv 1 \pmod{7}$. Thus $n_7 = 1$, so there is a unique Sylow 7-subgroup Q of G and $Q \triangleleft G$. Now $P \cap Q = \{1\}$ because $\gcd(|P|, |Q|) = 1$, so $|PQ| = |P||Q| = |G|$. Thus $G = PQ$, so $G \cong P \times Q$ by Theorem 3 §8.1. But P and Q are abelian by Theorem 7 §8.2, so G is abelian. \square

Example 7. Show that there is no simple group of order 56.

Solution. If $|G| = 56 = 2^3 \cdot 7$, then n_7 divides 8 and $n_7 \equiv 1 \pmod{7}$. This means that $n_7 = 1$ or $n_7 = 8$. If $n_7 = 1$, then the Sylow 7-subgroup is normal. If $n_7 = 8$, there are eight distinct cyclic subgroups in G of order 7. Because the intersection of any two of these subgroups equals $\{1\}$, there are $8 \cdot 6 = 48$ elements of order 7. This leaves eight elements, so the Sylow 2-subgroup is unique and hence normal. \square

Example 8. Show that there is no simple group of order 72.

Solution. If $|G| = 72 = 2^3 \cdot 3^2$, then $n_2 = 1, 3$, or 9 and $n_3 = 1$ or 4 by Theorem 3, so the method in Example 7 fails. However, let P denote any Sylow 3-subgroup of G . If $n_3 = 1$, then $P \triangleleft G$. If $n_3 = 4$, then $|G : N(P)| = 4$ by Sylow's third theorem. Thus, Theorem 1 §8.3 provides a homomorphism $\theta : G \rightarrow S_4$, with $\ker \theta \subseteq N(P)$, and $\ker \theta \neq \{1\}$ because $|G| = 72$ does not divide $|S_4| = 24$. As $\ker \theta \triangleleft G$ and $N(P) \neq G$, we are done. \square

A famous theorem of William Burnside asserts that no group of order $p^n q^m$ is simple where p and q are primes. The proof involves the theory of group representations and is beyond the scope of this book. However, we can do the following case.

Example 9. Show that no group of order $p^2 q^2$ is simple when p and q are primes.

Solution. Let $|G| = p^2 q^2$. If $p = q$, then $Z(G) \neq \{1\}$ by Theorem 6 §8.2 and we are finished. So assume that $p > q$. We have $n_p = 1, q$, or q^2 , and $n_p \equiv 1 \pmod{p}$. If $n_p = 1$, the Sylow p -subgroup is normal and we are done. The case $n_p = q$ is impossible because $q < p$. So assume that $n_p = q^2$, obtaining $q^2 \equiv 1 \pmod{p}$. Then p divides $q^2 - 1 = (q-1)(q+1)$, so either $p|(q-1)$ or $p|(q+1)$. Because $p > q$, the first alternative is impossible; the second implies that $q+1 \geq p > q$ from which $q+1 = p$. This means that $p = 3$ and $q = 2$, so $|G| = 36$. But then any Sylow 3-subgroup has index 4, so there is a homomorphism $\theta : G \rightarrow S_4$ by Theorem 1 §8.3. This homomorphism cannot be one-to-one, so $\ker \theta$ is normal in G . \square

We are now going to use the Sylow theorems to characterize the groups of order less than 16. It turns out that three of these groups belong to a family of groups that resemble the dihedral groups and are constructed in much the same way. We let $n = 2m$ be an even positive integer, let $w = e^{2\pi i/n}$, and let

$$A = \begin{bmatrix} w & 0 \\ 0 & w^{-1} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}.$$

Then A and B are invertible complex matrices, and we can easily verify that $|A| = n$, $ABA = B$, and $B^2 = A^m$. One verifies that

$$G = \{I, A, A^2, \dots, A^{n-1}, B, BA, \dots, BA^{n-1}\}$$

is a subgroup of $GL_2(\mathbb{C})$, and $|G| = 2n$ because $\langle A \rangle$ has index 2 in G . As for D_n , we abstract this situation as follows.

If $n = 2m$, $m \geq 1$, the **dicyclic group** Q_n is the group of order $2n$ presented as

$$Q_n = \{1, a, \dots, a^{n-1}, b, ba, \dots, ba^{n-1}\}, \quad \text{where } o(a) = n, aba = b, \text{ and } b^2 = a^m.$$

The condition that $|Q_n| = 2n$ amounts to the requirement that $b \notin \langle a \rangle$. The group Q_n is presented just like D_n , except that here $n = 2m$ must be even and $b^2 = a^m$ (recall that $b^2 = 1$ in the dihedral case). Again, $a^kba^k = b$ for all $k \in \mathbb{Z}$, so

$$a^k b = b a^{-k} = b a^{n-k}, \quad \text{for all } k \in \mathbb{Z}.$$

This equation shows that $(ba^k)^2 = b^2 = a^m$ for all k so, as $|a^m| = 2$, we get

$$|ba^k| = 4, \quad \text{for all } k \in \mathbb{Z}$$

(in contrast to $|ba^k| = 2$ in D_n). Two of these dicyclic groups are already familiar, as we see in Example 10.

Example 10. $Q_2 \cong C_4$ because $|Q_2| = 4$. We claim that $Q_4 \cong Q$, the quaternion group (Section 2.8). Indeed $Q = \{\pm 1, \pm i, \pm j, \pm k\}$ and, writing $a = i$ and $b = j$, we have $o(a) = 4$, $aba = b$, and $b^2 = a^2$. Hence $Q \cong Q_4$.

Theorem 4. Every group G of order 8 is isomorphic to one of C_8 , $C_4 \times C_2$, $C_2 \times C_2 \times C_2$, D_4 , or $Q_4 = Q$.

Proof. If G is abelian, then G is isomorphic to one of C_8 , $C_4 \times C_2$, or $C_2 \times C_2 \times C_2$ by Example 8 §2.8 (or by Corollary 2 of Theorem 6 §7.2). If G is not abelian, then $x^2 = 1$ cannot hold for all $x \in G$, so there exists $a \in G$, $o(a) = 4$. Write $K = \langle a \rangle$. If $b \notin K$, then $G = K \cup Kb$, and we claim that $aba = b$. Indeed $bab^{-1} \in K$ because $K \triangleleft G$, and $o(bab^{-1}) = o(a) = 4$. As $bab^{-1} \neq a$ because G is not abelian, we get $bab^{-1} = a^{-1}$; that is, $aba = b$. Hence, if $o(b) = 2$ for some $b \notin K$, then $G \cong D_4$. Otherwise, $o(b) = 4$ for all $b \notin K$. But then a^2 is the only element of G of order 2, so $b^2 = a^2$ for all $b \notin K$. Thus $G \cong Q_4$. ■

In order to determine the groups of order 12, we need Lemma 1.

Lemma 1. The only subgroup of S_n of index 2 is A_n .

Proof. If $|S_n : K| = 2$, then $K \triangleleft S_n$ and $|S_n/K| = 2$, so $\sigma^2 \in K$ for all $\sigma \in S_n$. If σ is a 3-cycle, then $\sigma^3 = \varepsilon$ so $\sigma = \sigma^4 \in K$. But A_n is generated by the 3-cycles (Lemma 4 §2.8), so $A_n \subseteq K$. This implies that $A_n = K$ because $|S_n : A_n| = 2$. ■

Theorem 5. Every group of order 12 is isomorphic to one of C_{12} , $C_6 \times C_2$, A_4 , D_6 , or Q_6 .

Proof. Let P and Q be Sylow subgroups with $|P| = 3$ and $|Q| = 4$. If G is abelian, $G \cong P \times Q$, so either $G \cong C_3 \times C_4 \cong C_{12}$ or $G \cong C_3 \times C_2 \times C_2 \cong C_6 \times C_2$. If G is nonabelian, there is a homomorphism $\theta : G \rightarrow S_4$ with $\ker \theta \subseteq P$. If $\ker \theta = \{1\}$, then $G \cong A_4$ by Lemma 2. So assume $P \triangleleft G$. Similarly, we have $\varphi : G \rightarrow S_3$ with $\ker \varphi \subseteq Q$. Write $\ker \varphi = L$. Then $L \neq \{1\}$, and $L = Q$ implies that $Q \triangleleft G$, so $G \cong P \times Q$ is abelian. Hence $|L| = 2$, so $LP \cong L \times P \cong C_6$.

So let $a \in G$ have order 6 and write $K = \langle a \rangle$. If $b \notin K$, then $G = K \cup Kb$ and $aba = b$, as in Theorem 4. Finally, $b^2 \in K$ because $|G/K| = 2$, and it remains to show that $b^2 = 1$ ($G \cong D_6$) or $b^2 = a^3$ ($G \cong Q_6$). If $b^2 = a$ or a^5 , then $o(b) = 12$ and G is abelian. If $b^2 = a^2$, then $b^3 = ba^2 = a^{-2}b = a^4b$, so $b^2 = a^4$, a contradiction. Hence $b^2 \neq a^2$ and, similarly, $b^2 \neq a^4$. ■

These results, together with earlier work, enable us to describe all the groups of order 15 or less. If p is a prime, the only group of order p is C_p . There are

two groups of order $2p$: C_{2p} and D_p (Theorem 3 §2.6). And there are two groups of order p^2 : C_{p^2} and $C_p \times C_p$ (Theorem 7 §8.2). We have already described the groups of order 8 or 12, and the only group of order 15 is C_{15} (Exercise 4). This list describes every group of order at most 15. The description of the groups of order 16 is more complicated (there are 14), and the general problem of describing all groups of a given order is extremely difficult.⁹³

We conclude this section with an elegant direct proof of Theorem 1. The argument requires a number-theoretic fact. Recall that the binomial coefficient $\binom{n}{r}$ is defined by $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ for $0 \leq r \leq n$.

Lemma 2. Let p be a prime and let m, n , and k be positive integers. Then p^n divides m if and only if p^n divides $\binom{p^k m}{p^k}$.

Proof. Since $\binom{p^k m}{p^k} = \frac{p^k m (p^k m - 1) \cdots (p^k m - i) \cdots [p^k m - (p^k - 1)]}{p^k (p^k - 1) \cdots (p^k - i) \cdots [p^k - (p^k - 1)]}$, it suffices to show that

$$p^n \text{ divides } (p^k m - i) \Leftrightarrow p^n \text{ divides } (p^k - i) \text{ for each } i = 1, 2, \dots, p^k - 1.$$

Observe that if p^n divides $(p^k m - i)$ then $n < k$ (otherwise $p^k | i$). Hence, the proof is completed by the observation that $(p^k m - i) = (p^k m - p^k) + (p^k - i)$. ■

With this we can give Helmut Wielandt's elegant proof⁹⁴ of Theorem 1, which does not use induction or Cauchy's theorem (and so provides another proof of Cauchy's theorem).

Proof of Theorem 1. If p^k divides $|G|$, let $X = \{U \subseteq G \mid U \text{ a subset, } |U| = p^k\}$ and define an action of G on X by $a \cdot U = aU$ for all U in X and a in G . Given U in X , let $S(U) = \{a \in G \mid aU = U\}$ denote the stabilizer. Write $|G| = p^k m$ and write $m = p^r w$, where p does not divide w .

Claim. V in X exists such that p^{r+1} does not divide $|G : S(V)|$.

Proof. If not, p^{r+1} divides the order of every orbit in X (by Lemma 3 §8.3) and so divides $|X|$. But $|X| = \binom{p^k m}{p^k}$, which means that p^{r+1} divides m by Lemma 3. Hence $p|w$, a contradiction. This proves the Claim.

Now let V be as in the claim and write $S = S(V)$. We show that $|S| = p^k$, so S is the desired subgroup of G . Now p^{r+1} does not divide $\frac{|G|}{|S|} = \frac{p^{r+k} w}{|S|}$ by the Claim, from which p^k does divide $|S|$. In particular, $p^k \leq |S|$. But if $v \in V$, then $Sv \subseteq V$ by the definition of S , so $|S| = |Sv| \leq |V| = p^k$. Thus $|S| = p^k$, as required. ■

Peter Ludvig Mejdell Sylow (1832–1918) Sylow was born in Norway and spent most of his professional life as a high school teacher in Halden. Despite onerous teaching duties, he found time to study the works of Abel and, in 1862–1863, he gave lectures on Galois theory and permutation groups at Christiania University in Oslo. The Sylow theorems were published in 1872 for permutation groups (Georg Frobenius extended them to abstract groups in 1887). These theorems are among the most important results

⁹³A lot is known about the number of groups of order p^n for a prime p . For example there are 2328 groups of order 2^7 , and 9310 groups of order 3^7 . See O'Brien, I.A. and Vaughan-Lee, M.R. *Journal of Algebra* 292 (2005), 243–258, for more information.

⁹⁴Wielandt, H., Ein Beweis Für die Existenz der Sylowgruppen, *Archiv der Mathematik* 10 (1959), 401–402.

on finite groups. Sylow applied them to show that any equation whose Galois group has prime-power order is solvable in radicals.

In addition to his study of groups, Sylow spent eight years editing the works of Abel. After his retirement from teaching high school, he was appointed to a chair at Christiania University, a position he held for the rest of his life.

Exercises 8.4

1. Find all Sylow 3-subgroups of S_4 and show explicitly that all are conjugate.
2. Find all Sylow 2-subgroups of D_n , where n is odd, and show explicitly that all are conjugate.
3. If P is a Sylow p -subgroup of G , prove that it is the only Sylow p -subgroup of $N(P)$.
4. Show that every group of order 15 is cyclic.
5. Show that there is only one group of order 1001.
6. Show that there are exactly two groups of order 99.
7. Show that a group G is not simple if
 - (a) $|G| = 40$
 - (b) $|G| = 80$
 - (c) $|G| = 48$
 - (d) $|G| = 108$
8. Show that no group of order 520 is simple.
9. Show that G has a cyclic normal subgroup of index 2 if
 - (a) $|G| = 70$
 - (b) $|G| = 154$
 - (c) $|G| = 30$
10. Show that G has a cyclic normal subgroup of index 5 if
 - (a) $|G| = 385$
 - (b) $|G| = 455$
11. (a) Show that G has a cyclic normal subgroup of index 3 if $|G| = 105$.
 (b) Show that G has an abelian normal subgroup of index 4 if $|G| = 700$.
12. If $|G| = pq$, where $p < q$ are primes and p does not divide $q - 1$, show that G is cyclic.
13. If $|G| = p^n m$, where $n \geq 1$, p is a prime, and $p > m$, show that the Sylow p -subgroup of G is normal in G .
14. If $|G| = p^2 q$, where p and q are primes, show that G is not simple.
15. If P is a normal Sylow p -subgroup of a finite group G , show that P is fully invariant in G ; that is, $\alpha(P) \subseteq P$ for every homomorphism $\alpha : G \rightarrow G$.
16. Let $P \triangleleft H$ and $H \triangleleft G$. If P is a Sylow subgroup of G , show that $P \triangleleft G$.
17. If P is a Sylow p -subgroup of G , show that $[N(P)]/P$ has no element of p -power order except the unity.
18. If P is a Sylow p -subgroup of G , let $N(P) \subseteq H$, H a subgroup of G .
 - (a) Show that $N(H) = H$. [Hint: If $a \in N(H)$, show that $aPa^{-1} \subseteq H$ and use Sylow's second theorem.]
 - (b) Show that p does not divide $|G : H|$.
19. If $N(P) = P$ for some Sylow p -subgroup P of G , show that $N(Q) = Q$ for every Sylow p -subgroup Q of G .
20. Suppose that $N(P) = P$ for some Sylow p -subgroup of the finite group G . Show that G/G' is an (abelian) p -group. [Hint: If $q \neq p$ is a prime divisor of $|G/G'|$, use Theorems 4 §7.2 and 8 §8.2 to find a subgroup H/G' of index q in G/G' . If Q is a Sylow p -subgroup of H , show that $N(Q) = Q$ by Exercise 18 and apply Exercise 19.]
21. Let K denote the intersection of all the Sylow p -subgroups of a finite group G . Show that K is a normal p -subgroup of G that contains every normal p -subgroup of G .
22. If $n = 2m$ and $m \geq 2$, show that $Z(Q_n) = \{1, a^m\}$.

23. If $k|n$, $k \geq 4$, and k is even, show that Q_n has a subgroup isomorphic to Q_k .
 24. If $k|n$, k is even, and n/k is odd, show that $K \triangleleft Q_n$ exists such that $Q_n/K \cong Q_k$.
 25. Show that, if G is a nonabelian group and $1 < |G| < 60$, then G is not simple. (Of course, $|A_5| = 60$ and A_5 is simple).

8.5 SEMIDIRECT PRODUCTS

There is no doubt that forming direct products is an important method of constructing groups using smaller groups: For example, all finite abelian groups can be constructed by forming direct products of cyclic groups. In this brief section, we describe a more general way to form a group from smaller groups.

Let K and H be subgroups of a group G and assume that $G = KH$, where $K \cap H = \{1\}$. If both $K \triangleleft G$ and $H \triangleleft G$ then $G \cong K \times H$ by Theorem 3 §8.1. In view of this, a natural question is what happens if only one of K and H is normal in G , say $K \triangleleft G$. Since $G = KH$ and $K \cap H = \{1\}$, each element $g \in G$ has a unique representation as $g = kh$ where $k \in K$ and $h \in H$. Given another element $g_1 = k_1 h_1$ of G , the key to understanding the group G is to describe the product gg_1 in the same form. This can be accomplished as follows:

$$gg_1 = khk_1h_1 = [k(hk_1h^{-1})](hh_1), \quad (*)$$

where $hk_1h^{-1} \in K$ because $K \triangleleft G$.

To describe this more formally, let $a \in G$ and define a map $\sigma_a : K \rightarrow K$ by $\sigma_a(k) = aka^{-1}$ for all $k \in K$. This makes sense because $K \triangleleft G$, and σ_a is an automorphism of K for each $a \in G$. Hence $(*)$ becomes

$$(kh)(k_1h_1) = [k\sigma_h(k_1)](hh_1). \quad (**)$$

Now observe that the map $\theta : H \rightarrow \text{aut}(K)$ given by $\theta(h) = \sigma_h$ is a group homomorphism.⁹⁵ This provides a way to turn this around: Starting with K , H , and θ , we can recreate G , using $(**)$ to motivate the multiplication.

Theorem 1. Let K and H be groups and let $\theta : H \rightarrow \text{aut}(K)$ be a homomorphism. Write $\theta(h) = \sigma_h$ for all $h \in H$. If $G = K \times H$ is the cartesian product, define an operation on G as follows:

$$(k, h)(k_1, h_1) = (k\sigma_h(k_1), hh_1), \quad \text{for all } (k, h) \text{ and } (k_1, h_1) \text{ in } G.$$

Write $K_1 = K \times 1$ and $H_1 = 1 \times H$. Then:

- (1) G is a group using this operation with unity $(1_K, 1_H)$ and inverses given by $(k, h)^{-1} = (\sigma_{h^{-1}}(k^{-1}), h^{-1})$.
- (2) Then K_1 and H_1 are subgroups of G , with $K_1 \cong K$ and $H_1 \cong H$.
- (3) $G = K_1H_1$, $K_1 \triangleleft G$ and $K_1 \cap H_1 = \{1\}$.

Proof. (2) and (3) are routine verifications. As to (1), the operation is associative because the products $[(k, h)(k_1, h_1)](k_2, h_2)$ and $(k, h)[(k_1, h_1)(k_2, h_2)]$ each simplify to $(k\sigma_h(k_1)\sigma_{h_1}(k_2), hh_1h_2)$. The verification of this, and of the rest of (1), is left to the reader. ■

⁹⁵This amounts to saying that $a \cdot k = \sigma_a(k)$ is an *action* of the group G on K . Such actions were studied in Section 8.3, where K was only required to be a set.

If K and H are groups and $\theta : H \rightarrow \text{aut } K$ is a homomorphism, the group G constructed in Theorem 1 is called the **semidirect product** of K by H , and is denoted $K \times_{\theta} H$. Theorem 1, and the discussion preceding it, give an important characterization of semidirect products (often taken as the definition).

Theorem 2. Let G be a group.

- (1) G is a semidirect product if and only if it has subgroups K and H with $G = KH$, $K \triangleleft G$ and $K \cap H = \{1\}$.
- (2) In that case, $G \cong K \times_{\theta} H$ for some homomorphism $\theta : H \rightarrow \text{aut } K$. Indeed, $\theta(h)$ is (the restriction to K of) conjugation by h for each $h \in H$.

Example 1. A direct product $K \times H$ is a semidirect product (let $\theta : H \rightarrow \text{aut } K$ be the trivial homomorphism: $\theta(h) = 1_K$ for each $h \in H$).

Example 2. The dihedral group $D_n = \{1, a, \dots, a^{n-1}, b, ba, \dots, ba^{n-1}\}$ is given by $o(a) = n$, $o(b) = 2$ and $aba = b$. If we write $K = \langle a \rangle$ and $H = \langle b \rangle$ then it is clear that $D_n = KH$, $K \triangleleft D_n$ (it has index 2) and $K \cap H = \{1\}$, so

$$D_n \cong C_n \times_{\theta} C_2 \quad \text{for some } \theta : C_2 \rightarrow \text{aut } C_n.$$

In fact, the multiplication in D_n is given by $a^k b = ba^{-k}$ for all k , so $ba^k b^{-1} = (a^k)^{-1}$. Hence $\theta : C_2 \rightarrow \text{aut } C_n$, where $\theta(b)$ is the automorphism $x \mapsto x^{-1}$.

It is interesting to observe that $D_6 \cong C_3 \times_{\theta} C_2$ by Example 2, and $C_6 \cong C_3 \times_{\theta} C_2$ by Example 1. Hence, a semidirect product $K \times_{\theta} H$ is *not* uniquely determined by K and H ; the homomorphism θ must be specified.

The next result determines all groups of order pq where p and q are primes. The theorem illustrates the way that semidirect products can be used to give the detailed structure of all groups of a given order, and it extends the theorem (Theorem 3 §2.6) that every group of order $2p$ is either cyclic or dihedral.

Theorem 3. Let G be a group of order pq where $p \leq q$ are primes.

- (1) If $p = q$ then G is cyclic or $G \cong C_p \times C_p$.
- (2) If $p < q$ and $q \not\equiv 1 \pmod{p}$ then G is cyclic.
- (3) If $p < q$ and $q \equiv 1 \pmod{p}$ then either G is cyclic or $G = \langle a, b \rangle$ where $o(a) = q$, $o(b) = p$ and $ab = ba^m$, and where $1 \leq m \leq q - 1$ and $m^p \equiv 1 \pmod{q}$. [Here, all choices for m result in isomorphic groups.]

Proof. By Cauchy's theorem, choose $a, b \in G$ such that $o(a) = q$ and $o(b) = p$, and write $K = \langle a \rangle$ and $H = \langle b \rangle$. Then $K \triangleleft G$ by the Corollary to Theorem 1 §8.3. Clearly $K \cap H = \{1\}$, so $G = KH$ and it follows that G is a semidirect product by Theorem 2.

- (1) This is Theorem 7 §8.2.
- (2) As usual, let n_p denote the number of Sylow p -subgroups of G . Then Sylow's third theorem (Theorem 3 §8.4) gives $n_p \equiv 1 \pmod{p}$ and $n_p \mid q$. Hence if $q \not\equiv 1 \pmod{p}$ then $n_p = 1$, so $H \triangleleft G$, and we have $G \cong K \times H \cong C_q \times C_p \cong C_{pq}$.
- (3) So assume $q \equiv 1 \pmod{p}$. Since $K = \langle a \rangle \triangleleft G$ let

$$b^{-1}ab = a^x \text{ where } x \in \mathbb{Z}. \tag{*}$$

Then $b^{-2}ab^2 = b^{-1}(a^x)b = (b^{-1}ab)^x = a^{x^2}$. Continuing in this way gives

$$b^{-k}ab^k = a^{x^k}, \quad \text{for any } k \geq 1.$$

But $b^p = 1$ so this gives $a^{x^p} = b^{-p}ab^p = a$, and hence $x^p \equiv 1 \pmod{q}$. Since $p \mid (q-1) = |\mathbb{Z}_q^*|$, let $\bar{m} \in \mathbb{Z}_q^*$ have order p (by Cauchy's theorem). Then $x = 1, m, m^2, \dots, m^{p-1}$ are all solutions to $x^p \equiv 1 \pmod{q}$. Moreover, there are no other solutions because $x^p - 1 = 0$ has at most p roots in \mathbb{Z}_q . If $x = 1$ in (*) then $ba = ab$ so G is abelian, and hence cyclic. If $x = m$ in (*) then $b^{-1}ab = a^m$ so $ab = ba^m$ and we have the situation in (3).

Finally, to realize the other solutions $x = m^r$ where $1 < r \leq p-1$, we change the generators of G . If we put $b_1 = b^r$ then $o(b_1) = p$ and $H = \langle b_1 \rangle$ because $o(b) = p$ is a prime. Hence $G = \langle a, b_1 \rangle$. Furthermore, $b_1^{-1}ab_1 = b^{-r}ab^r = a^{x^r} = a^{m^r}$, so this construction realizes the solution when $x = m^r$ in (*). ■

Other such theorems are possible, but we conclude by identifying one quite general situation in which a finite group G is necessarily a semidirect product, namely if G has a normal subgroup K and $|K|$ and $|G/K|$ are relatively prime. The next result of Issai Schur deals with the case when K is abelian.

Theorem 4. Schur's Theorem. Suppose G has an abelian normal subgroup K , where $|K|$ and $|G/K|$ are relatively prime. Then G has a subgroup of order $|G/K|$.

Proof. Put $|K| = m$ and $|G/K| = n$. In each coset a of G/K select an element $g_a \in G$ and assume that $g_1 = 1$, where 1 denotes the unity of G/K . If $a, b \in G/K$ then $g_a g_b = g_{ab} k(a, b)$ for a uniquely determined element $k(a, b) \in K$. If $g_a(g_b g_c) = (g_a g_b) g_c$ is written out it follows that

$$k(a, bc) k(b, c) = k(ab, c) [g_c^{-1} k(a, b) g_c]. \quad (*)$$

Now write $k(b) = \prod_{a \in G/K} k(a, b)$. If the product of each side of equation (*) is taken as a ranges over G/K , we obtain (since K is abelian)

$$k(bc)[k(b, c)]^n = k(c)[g_c^{-1} k(b) g_c]. \quad (**)$$

Now let $nn' \equiv 1 \pmod{m}$ and write $k(a)^{-n'} = k_a$ for all $a \in G/K$. Raising both sides of (**) to the power $-n'$ yields

$$k_{bc} k(b, c)^{-1} = k_c [g_c^{-1} k_{bc} g_c]. \quad (***)$$

Finally, write $H = \{g_a k_a \mid a \in G/K\}$. Then (***)) gives

$$\begin{aligned} (g_b k_b)(g_c k_c) &= [g_b g_c][g_c^{-1} k_b g_c] k_c \\ &= [g_{bc} k(b, c)][k_c^{-1} k_{bc} k(b, c)^{-1}] k_c \\ &= g_{bc} k_{bc}. \end{aligned}$$

This means that H is a subgroup and that the map $a \mapsto g_a k_a$ is an onto homomorphism $G/K \rightarrow H$. It is one-to-one because $g_a k_a = g_b k_b$ implies that $g_b^{-1} g_a = k_b k_a^{-1} \in K$, so $g_a = g_b$ by the choice of these elements. ■

Along with I. Schur, H. Zassenhaus is also credited with the general version of the next theorem.⁹⁶

⁹⁶The result was credited to Schur by Zassenhaus in his book *The Theory of Groups*, 2nd English Edition, Chelsea, 1958.

Theorem 5. Schur-Zassenhaus Theorem. Let G be a group of order kn where k and n are relatively prime. Assume that G has a normal subgroup K of order k . Then G has a subgroup H of order n , and so is a semidirect product $K \times_{\theta} H$.

Proof. It suffices to show that H exists (then $K \cap H = \{1\}$ because the orders are relatively prime, and hence $G = KH$). We may assume that $k > 1$. The proof is by induction on $|G|$.

Case 1. K contains a proper subgroup M such that $M \triangleleft G$.

Write $|M| = m$. Then $|G/M| = \frac{k}{m}n$, $K/M \triangleleft G/M$, and $|K/M| = \frac{k}{m}$, so G/M has a subgroup L/M of order n (by induction). But then $|L| = mn$, $M \triangleleft L$ and $|M| = m$, so L contains a subgroup of order n (again by induction). Hence, the theorem is proved in this case.

Case 2. K contains no proper subgroup that is normal in G .

In this case let P be a Sylow p -subgroup of K , and let $N = N(P)$ be the normalizer of P in G . Then P is a Sylow p -subgroup of G (because $\gcd(k, n) = 1$), and so has $|G : N|$ conjugates in G . These are all in K so, as $N \cap K$ is the normalizer of P in K , we obtain $|K : (N \cap K)| = |G : N|$. Hence $\frac{|K|}{|N \cap K|} = \frac{|G|}{|N|}$, so $|G| = \frac{|N||K|}{|K \cap N|} = |NK|$. Consequently $NK = G$ and so $N/(N \cap K) \cong G/K$ has order n .

If it happens that $N \neq G$ this shows that N contains a subgroup of order n (by induction), and we are done. So assume that $N = G$. This means that $P \triangleleft G$, and hence that $Z \triangleleft G$ where Z is the center of P (being characteristic in P , see Corollary 3 of Theorem 3 §2.8). Since $Z \neq 1$ by Theorem 6 §8.2, it follows that $Z = K$ (we are in Case 2). Thus, K is abelian and we are done by Schur's theorem (Theorem 4). ■

Exercises 8.5

1. (a) Show that $S_n \cong A_n \times_{\theta} C_2$ for some θ .
 (b) Show that the following is false: If K is a maximal subgroup that is normal, in G , then $G \cong K \times_{\theta} H$ for some H and θ . [Hint: The quaternion group.]
2. Find all groups of order 55.
3. Find all groups of order 39.
4. Show that there are two nonisomorphic groups of order 105. [Hint: Exercise 11(a) §8.4.]
5. Show that there are four nonisomorphic groups of order 30: C_{30} , D_{15} , $D_5 \times C_3$, and $D_3 \times C_5$. [Hint: Find a normal subgroup of index 2.]
6. Let $\alpha : G \rightarrow H$ be a group homomorphism, and write $\ker(\alpha) = K$. If there exists $\beta : H \rightarrow G$ such that $\alpha\beta = 1_H$, show that $G \cong K \times_{\theta} H$ for some θ .

8.6 AN APPLICATION TO COMBINATORICS

A main theme in this chapter has been to apply *counting* arguments to gain information about a finite group G by defining G -sets and using the orbit decomposition theorem. In this section, we turn this successful technique around and use the group to gain information about the sets it acts on. Specifically, we get a formula for the

number of distinct orbits, which is useful in solving certain combinatorial problems. We begin by deriving this formula that is of interest in its own right, and then describing how it applies to combinatorics.

If X is any G -set and $x \in X$, the stabilizer of x in X is the subgroup

$$S(x) = \{a \in G \mid a \cdot x = x\}$$

of all elements of G that fix x . Dually, if $a \in G$ we write

$$F(a) = \{x \in X \mid a \cdot x = x\},$$

the set of elements of X fixed by a . We refer to these sets frequently because of the following result of Cauchy and Frobenius.

Theorem 1. Cauchy–Frobenius Lemma.⁹⁷ *Let X be a G -set and assume that G and X are finite. If n is the number of distinct orbits of G in X , then*

$$n = \frac{1}{|G|} \sum_{a \in G} |F(a)|.$$

Proof. The proof proceeds by the time-honored method of counting the elements of a set Y in two ways and equating the results. In this case, consider the subset $Y = \{(a, x) \mid x \in X, a \in G, a \cdot x = x\}$ of $G \times X$. Then

$$|Y| = \sum_{a \in G} |F(a)| \tag{*}$$

because, for each element a of G , there are exactly $|F(a)|$ pairs in Y with first component a . In the same way, we obtain

$$|Y| = \sum_{x \in X} |S(x)|.$$

However, we can refine this second sum because X is partitioned into orbits by the action of G . If $G \cdot x_1, \dots, G \cdot x_n$ are the n distinct orbits, then each $x \in X$ belongs to exactly one orbit $G \cdot x_i$, so

$$|Y| = \sum_{i=1}^n \left[\sum_{x \in G \cdot x_i} |S(x)| \right]. \tag{**}$$

Now recall that $|G \cdot x| = |G : S(x)|$ holds for all $x \in X$ (Lemma 3 §8.3). If $x \in G \cdot x_i$, then $G \cdot x = G \cdot x_i$ and so $|S(x)| = |S(x_i)|$. Hence (**) becomes

$$|Y| = \sum_{i=1}^n \left[\sum_{x \in G \cdot x_i} |S(x_i)| \right] = \sum_{i=1}^n [|G \cdot x_i| |S(x_i)|] = \sum_{i=1}^n |G| = n|G|.$$

Combining this with (*) gives $n|G| = |Y| = \sum_{a \in G} |F(a)|$. The lemma follows. ■

As an illustration, let G act on itself by conjugation, so the orbits are the conjugacy classes. If $a \in G$, then $F(a) = \{x \in G \mid axa^{-1} = x\} = N(a)$ is the normalizer of a in G . Thus, the Cauchy–Frobenius lemma gives the following Corollary.

Corollary. *A finite group G has $(1/|G|) \sum_{a \in G} |N(a)|$ distinct conjugacy classes.*

Before applying the Cauchy–Frobenius lemma, we must consider a technicality. If G is a group and X is a nonempty set, a function $X \times G \rightarrow X$, written $(x, g) \mapsto x \cdot g$,

⁹⁷The lemma was known to Cauchy and Frobenius in the mid-1800s, and was rediscovered by William Burnside in 1900. Hence, it is also called Burnside's lemma.

is called a **right action** of G on X if $x \cdot 1 = x$ and $(x \cdot a) \cdot b = x \cdot (ab)$ hold for all $a, b \in G$ and all $x \in X$. Clearly, all the results for G -sets can be proved for right actions. In fact, we can easily verify that $a * x = x \cdot a^{-1}$ defines a (left) action of F on X . The reason for mentioning this is that right actions occur naturally in the examples that follow.

The combinatorial applications in which we are interested can all be described using the following format. Let D and C be nonempty finite sets and let C^D denote⁹⁸ the set of all mappings $\lambda : D \rightarrow C$. Suppose that G is a subgroup of S_D . Given $\sigma \in G$ and $\lambda \in C^D$, we have $D \xrightarrow{\sigma} D \xrightarrow{\lambda} C$, so it is natural to define

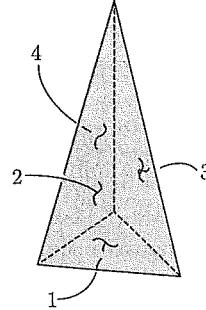
$$\lambda \cdot \sigma = \lambda \sigma = \text{the composition of the maps.}$$

This is a right action of G on the permutation group C^D (the axioms are elementary properties of composition of mappings), and it plays a central role in our discussion. Example 1 is typical.

Example 1. If q colors are available, find the number of ways in which a pyramid can be painted if the edges of the base are all of length 1 and the sides are of length 2. Assume that each face is painted a single color.

Solution. Label the faces 1, 2, 3, and 4 as shown in the figure at the right. Then the labeled pyramid can be colored in q^4 ways because there are q color choices for each face. The problem is that many of these colorings are indistinguishable when the labels are removed. The reason is that one labeled coloring may be carried to another by a motion of the pyramid, so both result in the same unlabeled coloring. To make this more precise, let

$$D = \{1, 2, 3, 4\} \quad \text{and} \quad C = \text{the set of } q \text{ colors.}$$



Then each map $\lambda : D \rightarrow C$ determines a labeled coloring, the color of face i being $\lambda(i)$. Conversely, each labeled coloring determines such a map, so we may identify C^D with the set of labeled colorings. Now let $G \subseteq S_D = S_4$ be the group of motions of the pyramid, where a motion is identified with the permutation of the face labels that it induces. Then G acts on C^D on the right as discussed previously, and we claim that the unlabeled colorings can be identified with the orbits of G in the set C^D of labeled colorings. Indeed, if λ and μ are labeled colorings in C^D , then

- λ and μ lead to indistinguishable colorings when the labels are removed;
- $\Leftrightarrow \lambda$ is achieved by first moving the pyramid and then applying μ ;
- $\Leftrightarrow \lambda = \mu\sigma$ for some $\sigma \in G$;
- $\Leftrightarrow \lambda$ and μ are in the same G -orbit.

Hence, the number of unlabeled colorings is equal to the number of orbits, so the Cauchy–Frobenius lemma applies. In this case $G = \{\varepsilon, \sigma, \sigma^2\}$, where $\sigma = (2 \ 3 \ 4)$. We have $F(\varepsilon) = C^D$, so $|F(\varepsilon)| = q^4$. Next,

$$F(\sigma) = \{\lambda \mid \lambda\sigma = \lambda\} = \{\lambda \mid \lambda(2) = \lambda(3) = \lambda(4)\}.$$

⁹⁸This exponential notation is used because $|C^D| = |C|^{|D|}$.

Hence a coloring λ is in $F(\sigma)$ just when sides 2, 3, and 4 are all the same color. We may choose this color in q ways and color the base in q ways, so $|F(\sigma)| = q^2$. Similarly, $|F(\sigma^2)| = q^2$, so the number of orbits is $\frac{1}{3}(q^4 + 2q^2)$ by Theorem 1. \square

The technique used in Example 1 can be used in the same way to count the number of ways to color the edges or vertices of a figure. In general, we label the objects to be colored as $1, 2, 3, \dots, n$. The group G is the subgroup of S_n consisting of all permutations of these objects resulting from a rigid motion of the figure. We then identify the colorings with the set C^D of all mappings from $D = \{1, 2, \dots, n\}$ to the set C of colors. If λ is such a map and $\sigma \in G$, the map $\lambda\sigma$ colors object i the same as the map λ colors object $\sigma(i)$. As σ is a motion of the figure, the results are indistinguishable when the labels are removed, so the number of distinguishable unlabeled colorings equals the number of orbits (as in Example 1). Hence, the Cauchy–Frobenius lemma applies.

Before giving more examples, we describe a convenient way to compute $|F(\sigma)|$ in the Cauchy–Frobenius lemma, where $\sigma \in S_n$, $D = \{1, 2, \dots, n\}$, and S_n acts on C^D , as before. If σ is factored into disjoint cycles, we customarily ignore a cycle (k) of length 1 because σ fixes k . However, our present purpose requires that we include such cycles.

For example, if $n = 7$, we now think of $\sigma = (1 \ 4)(3 \ 5 \ 7)$ in S_7 as a product of four disjoint cycles: $\sigma = (1 \ 4)(2)(3 \ 5 \ 7)(6)$. If q colors are available in C , we claim that $|F(\sigma)| = q^4$. Indeed, given $\lambda : D \rightarrow C$, we have

$$\lambda \in F(\sigma) \Leftrightarrow \lambda\sigma = \lambda \Leftrightarrow \lambda(1) = \lambda(4) \text{ and } \lambda(3) = \lambda(5) = \lambda(7),$$

so there are q choices for each of the colors $\lambda(1) = \lambda(4)$, $\lambda(2)$, $\lambda(3) = \lambda(5) = \lambda(7)$, and $\lambda(6)$ and hence q^4 possibilities for the map λ .

The obvious generalization is valid. If $\sigma \in S_n$, then $|F(\sigma)| = q^c$, where c is the number of cycles in the factorization in S_n of σ into disjoint cycles (including cycles of length 1). The integer c is called the **cycle index** of σ and is denoted $c = \text{cyc } \sigma$. We record this as Theorem 2.

Theorem 2. Let C be a set of q colors and let S_n act on C^D by composition of maps, where $D = \{1, 2, \dots, n\}$. Then

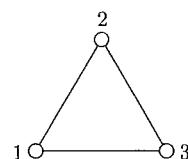
$$|F(\sigma)| = q^{\text{cyc } \sigma} \text{ for any } \sigma \in S_n.$$

If G is a subgroup of S_n , the number of orbits of G in C^D is

$$(1/|G|) \sum_{\sigma \in G} q^{\text{cyc } \sigma}.$$

Example 2. Suppose that a chemical molecule is modeled in the form of an equilateral triangle with the atoms at the vertices as shown in the figure below. If q colors are available and each atom is painted a single color, how many distinct ways can the molecule be colored? (The edges are not painted.)

Solution. Here the three vertices, labeled 1, 2, and 3, are permuted by motions in S_3 . Because of the high degree of symmetry of the equilateral triangle, every permutation in S_3 can be achieved by a motion, so S_3 is the group of motions. By Theorem 2 we get



$$\begin{aligned} |F(\varepsilon)| &= q^3, \\ |F(1 \ 2 \ 3)| &= |F(1 \ 3 \ 2)| = q, \\ |F(1 \ 2)| &= |F(1 \ 3)| = |F(2 \ 3)| = q^2. \end{aligned}$$

By Theorem 1 there are $\frac{1}{6}(q^3 + 2q + 3q^2) = \frac{1}{6}q(q+1)(q+2)$ colorings (orbits). \square

In Example 3, we vary the theme by insisting that no color is repeated. This amounts to labeling the various facets of the object with distinct colors.

Example 3. Suppose that children's blocks are to be constructed as cubes with each of the six faces painted a different color. If $q \geq 6$ colors are available, how many distinct blocks can be made?

Solution. Let $D = \{1, 2, 3, 4, 5, 6\}$ and let C be the set of q colors, as before. Because the faces are distinct colors, a coloring in this case is a one-to-one mapping $\lambda : D \rightarrow C$. Let $X \subseteq C^D$ denote the set of all such mappings. If G is the group of motions of the cube, G acts on X by composition because $\lambda\sigma$ is one-to-one whenever $\sigma \in G$ and $\lambda \in X$. If $\sigma \neq \varepsilon$ in G , then $F(\sigma) = \{\lambda \mid \lambda\sigma = \lambda\}$ is empty. (If $\lambda \in F(\sigma)$, then $\sigma(i) = j$ implies that $\lambda(j) = \lambda[\sigma(i)] = \lambda(i)$, from which $i = j$). Thus $|F(\sigma)| = 0$ if $\sigma \neq \varepsilon$, whereas $|F(\varepsilon)| = q!/(q-6)!$ because $F(\varepsilon) = X$. Hence, the Cauchy–Frobenius lemma gives the number of colorings as $q!/(|G|(q-6)!)$, so it remains only to compute $|G|$. Label the faces of the cube 1, 2, 3, 4, 5, and 6. If we initially place the cube with side 1 on top, we determine a motion by choosing which side ends up on top (six choices) and then choosing one of four rotations fixing the top and bottom faces (four choices). Thus, there are $6 \cdot 4 = 24$ choices in all, so $|G| = 24$ and there are $q!/(24(q-6)!)$ possible blocks. If $q = 6$ (the minimal number of colors), there are $6!/(24 \cdot 24) = 30$ possible blocks. \square

The argument in Example 3 gives the general result in Theorem 3.

Theorem 3. Let $D = \{1, 2, \dots, n\}$, let C be a set of $q \geq n$ colors, and let $X \subseteq C^D$ denote the set of one-to-one mappings $D \rightarrow C$. If G is a subgroup of S_n , then G acts on X by composition of maps, and the number of orbits is $q!/(|G|(q-n)!)$.

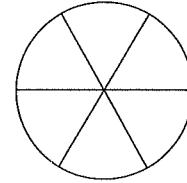
Needless to say, this theory has been developed further, and these examples provide only a glimpse of the possibilities. For example, we could ask how many ways a cube can be painted with q colors when exactly two faces are red or when at least two faces are red. In 1937, George Polya answered such questions, and many others, by giving an elegant and comprehensive generalization of the Cauchy–Frobenius lemma.⁹⁹ This is beyond the scope of this book.

Exercises 8.6

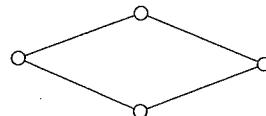
1. If H is a subgroup of a finite group G , use the Cauchy–Frobenius lemma to compute the number of distinct right cosets of H in G .
2. Verify the Corollary to the Cauchy–Frobenius lemma when $G = S_3$.
3. (a) If q colors are available, show that there are $\frac{1}{2}q^2(q+1)$ ways to paint the vertices of an isosceles triangle (not equilateral).
 (b) Derive the formula in (a) by using elementary counting methods.

⁹⁹For an exposition of Polya's theory (by N. G. de Bruijn), see Beckenbach, E., ed., *Applied Combinatorial Mathematics*, New York: Wiley, 1964. Another good treatment appears in Roberts, F.S., *Applied Combinatorics*, Englewood Cliffs, NJ: Prentice-Hall, 1984, Chapter 7.

4. (a) If q colors are available, show that there are $\frac{1}{12}q^2(q^2 + 11)$ ways to paint the faces of a tetrahedron (four faces, each an equilateral triangle). [Hint: Example 3 §2.7.]
 (b) Repeat (a) if $q \geq 4$ and no two faces are the same color.
5. (a) If q colors are available, show that there are $\frac{1}{8}q^3(q + 1)(q^2 - q + 4)$ ways to paint the faces of a rectangular solid with square ends (not a cube).
 (b) Repeat (a) if $q \geq 6$ and no two faces are the same color.
6. If q colors are available, how many ways can
 (a) the vertices of a tetrahedron be painted?
 (b) the edges of a tetrahedron be painted?
7. How many ways can the faces of a cube be painted with q colors? [Hint: The group G of motions has $|G| = 24$. Here G consists of ε and various rotations: nine about a line through the centers of opposite faces, six about a line through the centers of opposite edges, and eight about a line through opposite vertices.]
8. (a) A circular disk is divided into six equal sections, as shown in the figure at the right. If q colors are available, how many ways can one side of the disk be painted if each section is painted a single color? How many if no two sections are the same color.
 (b) Repeat (a) if the sections are made of transparent glass and the circle can be turned over.
9. Show that there are $\frac{1}{2}[q^n + q^{\lfloor(n+1)/2\rfloor}]$ ways to make a rectangular necktie with n stripes if there are q colors. (Here $\lfloor k \rfloor$ denotes the greatest integer $\leq k$.)
10. Assume that q colors are available for painting the vertices and r colors are available for painting the edges of an equilateral triangle. Show that there are $\frac{1}{6}qr(qr + 1)(qr + 2)$ ways to paint both edges and vertices. [Hint: A motion σ of the triangle induces a permutation σ_v of the vertices and a permutation σ_e of the edges. Let σ act in the obvious way on pairs (λ, μ) , where λ and μ are vertex and edge colorings, respectively.]
11. Repeat Exercise 10 with a planar figure as shown at the right, where the four outer edges have the same length and the inner edge is shorter.



12. If G is a finite group, let $p(G)$ denote the probability that $ab = ba$, where a and b are selected at random (with replacement) from G .
 (a) Show that $p(G) = [k(G)]/|G|$, where $k(G)$ is the number of distinct conjugacy classes of G .
 (b) Show that $p(G) \leq \frac{5}{8}$ if G is nonabelian, with equality for a suitable group G of order 8.



Chapter 9

Series of Subgroups

In the future, as in the past, the great ideas must be simplifying ideas.

—André Weil

If G is a finite abelian group, it can be shown that G is isomorphic to a direct product of cyclic groups (see Chapter 7). This result is an example of a structure theorem, that is, a theorem showing that every group in a suitable defined class may be constructed in a systematic way from well understood groups in the class. Such theorems are hard to come by, and the result for finite abelian groups is a stunning example. The structure of nonabelian finite groups is much more complicated.

Suppose that groups K and H are given. It is a very difficult problem to describe all groups G that have a normal subgroup K_1 isomorphic to K such that G/K_1 is isomorphic to H . If we could solve this *extension problem*, the solution would give an inductive method for constructing all finite groups. Direct and semidirect products solve this problem in very special cases. Although the general problem is far from being solved, the classes of groups that can be built up this way are of interest.

To illustrate, suppose that we use only abelian groups as building blocks. Starting with an abelian group G_0 , we construct $G_1 \supseteq G_0$ such that $G_0 \triangleleft G_1$ and G_1/G_0 is abelian. Next, we extend G_1 to obtain $G_2 \supseteq G_1$ such that $G_1 \triangleleft G_2$ and G_2/G_1 is abelian. After n steps, we have a chain

$$G = G_n \supseteq G_{n-1} \supseteq \cdots \supseteq G_1 \supseteq G_0,$$

where $G_i \triangleleft G_{i+1}$ and G_{i+1}/G_i is abelian for each i . Such a group G is called *solvable*, and the theory of these groups is successful in the following sense: The class of solvable groups is large (it contains all finite groups of odd order), but at the same time, many theorems are true for all solvable groups but do not hold in general. We investigate solvable groups in Section 9.2.

If we use simple groups as building blocks in this way, the resulting groups are those studied in Section 9.1. In this case, the above chain of subgroups is called a *composition series* for G , and the famous Jordan–Hölder theorem asserts that G uniquely determines the series of groups G_n/G_{n-1} , G_{n-1}/G_{n-2} , ..., $G_1/G_0, G_0$. This leads to the useful notion of the composition length of a group.

Section 9.3 deals with somewhat more specialized central series and begins the study of finite nilpotent groups. These groups are characterized as the groups that are the direct product of their Sylow subgroups, equivalently if every Sylow subgroup is normal. In addition, the Frattini subgroup is defined for every finite group and shown to be nilpotent.

9.1 THE JORDAN–HÖLDER THEOREM

Much of what we do in this chapter is concerned with groups G that admit a chain of subgroups with certain nice properties. A **subnormal series** for G is a chain

$$G = G_0 \supseteq G_1 \supseteq G_2 \supseteq \cdots \supseteq G_n = \{1\}$$

of subgroups of G such that $G_{i+1} \triangleleft G_i$ for each i . The factor groups G_i/G_{i+1} are called the **factors** of the subnormal series. Note that we do not insist that the subgroups G_i are normal in G . Moreover, by possibly deleting some of the groups G_i , we clearly may assume that $G_i \neq G_{i+1}$ for each i .

Example 1. $G \supseteq \{1\}$ is a subnormal series for any group G . The only factor is G .

Example 2. $A_4 \supset K \supset H \supset \varepsilon$ is a subnormal series for A_4 , we have $K = \{\varepsilon, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$ and $H = \{\varepsilon, (1\ 2)(3\ 4)\}$. The factors are C_3, C_2 , and C_2 in order. Note that H is not normal in A_4 .

The most interesting cases are the groups that admit a subnormal series in which every factor is abelian or every factor is simple. We investigate the first case in Section 9.2; the second calls for another definition.

If G is a group, a subnormal series $G = G_0 \supset G_1 \supset \cdots \supset G_n = \{1\}$ is called a **composition series** for G if each factor G_i/G_{i+1} is simple. In this case, the factors G_i/G_{i+1} are called **composition factors** of G , and the integer n is called the **length** of the composition series. If $G = \{1\}$, we say that G has a composition series of length 0.

Example 3. The simple groups are those with a composition series of length 1.

Example 4. Every finite group G has a composition series. This holds by definition if $G = \{1\}$. If $G \neq \{1\}$, write $G = G_0$ and choose a maximal normal subgroup G_1 of G_0 (it exists because G is finite). Then G_0/G_1 is simple by Theorem 6 §8.1. If $G_1 \neq \{1\}$, choose a maximal normal subgroup G_2 of G_1 and continue in this way. The series $G = G_0 \supset G_1 \supset G_2 \supset \cdots$ must reach $\{1\}$ eventually because G is finite, so it is a composition series.

The converse of Example 4 is false: Any infinite simple group has a composition series of length 1. However, the converse does hold for abelian groups.

Example 5. An abelian group G has a composition series if and only if G is finite. Indeed, if $G = G_0 \supset G_1 \supset \cdots \supset G_n = \{1\}$ is a composition series, each composition

factor G_i/G_{i+1} is a simple abelian group and so is finite. Hence (Exercise 11), $|G| = |G/G_1||G_1/G_2|\cdots|G_{n-1}/G_n|$ is also finite.

The finite abelian groups are not the only ones having all composition factors abelian. Theorem 8 §8.2 shows that every finite p -group has this property:

Example 6. If p is a prime, each finite p -group has a composition series in which every composition factor is isomorphic to C_p .

A group may have several different composition series. For example, let G be a cyclic group of order 12 and, for each divisor d of 12, let H_d be the unique subgroup of G of order d . Then G has three composition series. These series, along with their composition factors, are

$$\begin{aligned} G &\supset H_6 \supset H_3 \supset \{1\}, & \text{Factors: } C_2, C_2, C_3, \\ G &\supset H_6 \supset H_2 \supset \{1\}, & \text{Factors: } C_2, C_3, C_2, \\ G &\supset H_4 \supset H_2 \supset \{1\}, & \text{Factors: } C_3, C_2, C_2. \end{aligned}$$

Note that the length is 3 in each case and the factors are also the same except for the order in which they appear. Hence, these series are all equivalent in the following sense: Two composition series for a group are said to be **equivalent** if they have the same length and the composition factors can be paired in such a way that corresponding factors are isomorphic. This is clearly an equivalence relation on the set of all composition series of a group G (assuming that there is one). The remarkable thing is that *any* two composition series for G are equivalent. This is the most important theorem in this section.

Theorem 1. Jordan–Hölder Theorem. *If a group has a composition series, any two composition series are equivalent.*

We will give the proof at the end of this section.

If a group G has a composition series, it uniquely determines the length of the series and the composition factors (including multiplicities). Hence, we can speak of the **composition length** of the group G , denoted $\text{length } G$, and of the **composition factors** of G .

Composition series were first discussed by Camille Jordan. In 1869, he showed (for groups of permutations) that the orders of the composition factors are the same for every composition series of the group. However, it was not until 20 years later, after the abstract definition of a group had been given, that Otto Hölder observed that the group uniquely determines the composition factors themselves and that they are the same in any composition series.

Example 7. If $n \geq 5$, then $S_n \supset A_5 \supset \{\varepsilon\}$ is a composition series because A_n is simple (Theorem 8 §2.8). Hence, S_n has length 2 and the composition factors are $C_2 \cong S_n/A_n$ and A_n . If $n = 4$, we get a composition series

$$S_4 \supset A_4 \supset K \supset H \supset \{\varepsilon\},$$

where $K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$ and $H = \{\varepsilon, (1 \ 2)(3 \ 4)\}$. Hence, S_4 has length 4 and the composition factors are C_2, C_3, C_2 , and C_2 .

The Jordan–Hölder theorem is a type of unique factorization theorem. In the following corollary, we use it to give another proof that the factorization of an integer n into primes is unique. Here, the composition factors play the role of primes.

Corollary. *The factorization of an integer $n \geq 2$ into primes is unique.*

Proof. Let $n = p_1 p_2 \cdots p_r$, where p_i are (not necessarily distinct) primes. If $G = \langle g \rangle$ is a cyclic group of order n , then

$$G = \langle g \rangle \supset \langle g^{p_1} \rangle \supset \langle g^{p_1 p_2} \rangle \supset \cdots \supset \langle g^{p_1 p_2 \cdots p_r} \rangle = \{1\}$$

is a composition series for G because the factors are $C_{p_1}, C_{p_2}, \dots, C_{p_r}$. (Indeed, $o(g) = n$ and $o(g^{p_1 p_2 \cdots p_k}) = p_{k+1} \cdots p_r$ for each k by Theorem 5 §2.4.) Since any other factorization of n into primes must yield the same composition factors, it follows that n uniquely determines the number $r = \text{length } G$ and the primes p_i . ■

If $K \triangleleft G$ and $G/K \cong H$, the group G is called an **extension** of K by H . So if

$$G = G_0 \supset G_1 \supset \cdots \supset G_r = \{1\}$$

is a composition series for G , then each G_i is an extension of G_{i+1} by a simple group. Thus, each finite group G is the result of a finite number of extensions by finite simple groups, and the Jordan–Hölder theorem shows that G uniquely determines the simple groups used (up to order). Moreover, we know all the finite simple groups (see the discussion at the end of Section 2.8), so the complete description of all finite groups comes down to the **extension problem**: For a simple group H , describe all extensions of a given group G by H . This is a very difficult task.

We are going to prove that subgroups and homomorphic images of groups with a composition series again have composition series. The proof requires the following technical lemma that gives important information about subnormal series in general and will be referred to again later. For composition series, we use it to deduce some important properties of the length of a group and prove the Jordan–Hölder theorem itself.

Lemma 1. *Let $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ be a subnormal series for the group G and let $K \triangleleft G$.*

- (1) *$K = K \cap G_0 \supseteq K \cap G_1 \supseteq \cdots \supseteq K \cap G_n = \{1\}$ is a subnormal series for K , and the factor $(K \cap G_i)/(K \cap G_{i+1})$ is isomorphic to a normal subgroup of G_i/G_{i+1} for each i .*
- (2) *$G/K = (KG_0)/K \supseteq (KG_1)/K \supseteq \cdots \supseteq (KG_n)/K = \{K\}$ is a subnormal series for G/K , and the factor $[(KG_i)/K]/[(KG_{i+1})/K]$ is a homomorphic image of G_i/G_{i+1} for each i .*

Proof. We leave to the reader the verification that the series are subnormal.

(1) Define $\alpha : K \cap G_i \rightarrow G_i/G_{i+1}$ by $\alpha(x) = xG_{i+1}$. This is clearly a group homomorphism and $\ker \alpha = \{x \in K \cap G_i \mid x \in G_{i+1}\} = K \cap G_{i+1}$. Hence, it remains to prove that $\alpha(K \cap G_i) \triangleleft (G_i/G_{i+1})$. But if $x \in K \cap G_i$ and $y \in G_i$, then

$$(yG_{i+1})\alpha(x)(yG_{i+1})^{-1} = (yxy^{-1})G_{i+1} = \alpha(yxy^{-1}) \in \alpha(K \cap G_i)$$

because $yxy^{-1} \in (yKy^{-1}) \cap G_i = K \cap G_i$.

(2) Since $KG_{i+1} \triangleleft KG_i$ (Exercise 15 §8.1), the third isomorphism theorem (Theorem 7 §8.1) shows that

$$\frac{(KG_i)/K}{(KG_{i+1})/K} \cong \frac{KG_i}{KG_{i+1}}.$$

To show that $(KG_i)/(KG_{i+1})$ is an image of G_i/G_{i+1} , define

$$\alpha : \frac{G_i}{G_{i+1}} \rightarrow \frac{KG_i}{KG_{i+1}} \quad \text{by} \quad \alpha(xG_{i+1}) = xKG_{i+1} \text{ for all } x \in G_i.$$

This is well defined because $xG_{i+1} = yG_{i+1}$ implies that $y^{-1}x \in G_{i+1} \subseteq KG_{i+1}$. It is clearly a group homomorphism, and (as $K \triangleleft G$) it is onto because

$$kxKG_{i+1} = x(x^{-1}kx)KG_{i+1} = xKG_{i+1} = \alpha(xG_{i+1})$$

holds for all $k \in K$ and $x \in G_i$. ■

Now suppose that $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ is a composition series for G . If $K \triangleleft G$, the subnormal series for K and G/K in Lemma 1 are also composition series. Indeed, in both cases, the factors are isomorphic to either normal subgroups or images of the simple groups G_i/G_{i+1} and so are all either simple or $\{1\}$. Hence, after we eliminate equalities, these series become composition series for K and G/K , respectively, each with factors from G . This proves part of Theorem 2.

Theorem 2. *Let G be a group and let $K \triangleleft G$. Then G has a composition series if and only if both K and G/K have composition series. Moreover, in this case,*

- (1) $\text{length } G = \text{length } K + \text{length } G/K$.
- (2) *The composition factors of G are exactly those of K and G/K .*
- (3) *G has a composition series containing K .*

Proof. If G has a composition series, we have already seen that this is true of K and G/K . Conversely, suppose

$$K = K_0 \supset K_1 \supset \cdots \supset K_m = \{1\} \quad \text{and} \quad \frac{G}{K} = \frac{G_0}{K} \supset \frac{G_1}{K} \supset \cdots \supset \frac{G_r}{K} = \{K\}$$

are composition series for K and G/K . Because $\frac{G_i}{G_{i+1}} \cong \frac{G_i/K}{G_{i+1}/K}$, the series

$$G = G_0 \supset G_1 \supset \cdots \supset G_r = K = K_0 \supset K_1 \supset \cdots \supset K_m = \{1\}$$

is a composition series for G . Now (1)–(3) are apparent. ■

Corollary 1. *If $\theta : G \rightarrow H$ is a homomorphism, then G has a composition series if and only if both $\ker \theta$ and $\theta(G)$ have composition series. In this case,*

$$\text{length } G = \text{length } \ker \theta + \text{length } \theta(G).$$

Proof. Since $G/\ker \theta \cong \theta(G)$, Theorem 2 applies with $K = \ker \theta$. □

Corollary 2. *If G_1, G_2, \dots, G_n are groups, then $G_1 \times G_2 \times \cdots \times G_n$ has a composition series if and only if the same is true of each G_i . In this case,*

$$\text{length}(G_1 \times G_2 \times \cdots \times G_n) = \text{length } G_1 + \text{length } G_2 + \cdots + \text{length } G_n.$$

Solution. We prove it for $n = 2$, and then the general case follows by induction. Define $\theta : G_1 \times G_2 \rightarrow G_2$ by $\theta(g_1, g_2) = g_2$ for all $g_1 \in G_1$ and $g_2 \in G_2$. Then θ is a homomorphism, $\ker \theta = G_1 \times \{1\} \cong G_1$, and $\theta(G_1 \times G_2) = G_2$. Use Corollary 1. □

Example 8. Let G be an abelian group of order $p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, where p_i are distinct primes. Show that $\text{length } G = n_1 + n_2 + \cdots + n_r$.

Solution. Proceed by induction on $|G|$. If $|G| = 1$, it is clear. In general, let K be a maximal (normal) subgroup of G . Then $|G/K|$ is a prime divisor of $|G|$, say $|G/K| = p_1$. Hence, $|K| = p_1^{n_1-1} p_2^{n_2} \cdots p_r^{n_r}$, so induction and Theorem 2 give

$$\begin{aligned}\text{length } G &= \text{length } K + \text{length}(G/K) = [(n_1 - 1) + n_2 + \cdots + n_r] + 1 \\ &= n_1 + n_2 + \cdots + n_r.\end{aligned}$$

□

We conclude this section with a proof of the Jordan–Hölder theorem. The proof requires the following lemma.

Lemma 2. *Let G be a group and let H and K be distinct maximal normal subgroups of G . Then $H \cap K$ is maximal normal in both H and K . Moreover,*

$$\frac{H}{H \cap K} \cong \frac{G}{K} \quad \text{and} \quad \frac{K}{K \cap H} \cong \frac{G}{H}.$$

Proof. We first claim that $KH = G$. Indeed, $H \subseteq KH \triangleleft G$ and $K \subseteq KH \triangleleft G$, so if $KH \neq G$, the fact that H and K are maximal normal in G implies that $H = KH = K$, contrary to assumption. Hence, $KH = G$, so the second isomorphism theorem (Theorem 1 §8.1) gives $\frac{G}{K} = \frac{KH}{K} \cong \frac{H}{H \cap K}$. Because G/K is simple, this shows that $K \cap H$ is maximal normal in H . The rest is proved in the same way. ■

Proof of the Jordan–Hölder Theorem. Suppose the group G has a composition series

$$G = G_0 \supset G_1 \supset G_2 \supset \cdots \supset G_n = \{1\} \tag{1}$$

of length n . We show by induction on n that every composition series

$$G = H_0 \supset H_1 \supset H_2 \supset \cdots \supset H_m = \{1\} \tag{2}$$

for G is equivalent to series (1). If $n = 1$, then G is simple, so $G_1 = \{1\} = H_1$ and the theorem holds. So assume that $n \geq 2$ and that the theorem holds for all groups with a composition series of length less than n . In particular, it holds for G_1 because $G_1 \supset G_2 \supset \cdots \supset G_n = \{1\}$ has length $n - 1$. If it happens that $H_1 = G_1$, then $G_1 \supset H_2 \supset \cdots \supset H_m = \{1\}$ is another composition series for G_1 and so is equivalent to $G_1 \supset G_2 \supset \cdots \supset G_n = \{1\}$ by induction and the theorem follows.

So assume that $H_1 \neq G_1$ and let $H_1 \cap G_1 = L_0 \supset L_1 \supset \cdots \supset L_s = \{1\}$ be a composition series for $H_1 \cap G_1$ by Theorem 2 as $H_1 \cap G_1 \triangleleft G$. Now consider the following series for G :

$$G \supset G_1 \supset (H_1 \cap G_1) \supset L_1 \supset \cdots \supset L_s = \{1\}, \tag{3}$$

$$G \supset H_1 \supset (H_1 \cap G_1) \supset L_1 \supset \cdots \supset L_s = \{1\}. \tag{4}$$

As $H_1 \neq G_1$, Lemma 2 asserts that $H_1 \cap G_1$ is maximal normal in each of H_1 and G_1 , so both (3) and (4) are composition series for G . Moreover, $G/G_1 \cong H_1/(H_1 \cap G_1)$ and $G/H_1 \cong G_1/(H_1 \cap G_1)$ by Lemma 2, so (3) and (4) are equivalent; denote this as (3) ~ (4). Note that this equivalence holds even if $s = 0$, that is, if $H_1 \cap G_1 = \{1\}$.

Now $G_1 \supset G_2 \supset \cdots \supset G_n = \{1\}$ and $G_1 \supset (H_1 \cap G_1) \supset L_1 \supset \cdots \supset L_s = \{1\}$ are composition series for G_1 and so are equivalent by induction. This implies that (1) ~ (3) and also that $n - 1 = s + 1$. But then the composition series $H_1 \supset (H_1 \cap G_1) \supset L_1 \supset \cdots \supset L_s = \{1\}$ has length $n - 1$ and so (again by induction) is equivalent to $H_1 \supset H_2 \supset \cdots \supset H_m = \{1\}$. This in turn implies that (2) ~ (4). Piecing these equivalent series together gives (1) ~ (3) ~ (4) ~ (2), which proves the Jordan–Hölder theorem. ■

Exercises 9.1

1. In each case, find the length of the group and exhibit the composition factors.
 - (a) C_8
 - (b) C_{12}
 - (c) D_4
 - (d) A_4
 - (e) Q (see Section 2.8)
2. If $n \geq 1$ and p is prime, show that C_{p^n} has exactly one composition series.
3. Find all composition series for
 - (a) C_{24}
 - (b) C_{30}
4. Find two nonisomorphic finite groups with identical composition factors.
5. Find all composition series for $C_4 \times C_2$.
6. If $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, find the length of D_n . [Hint: Example 8.]
7. Find a composition series for D_{16} that contains the center $Z(D_{16})$ and find one that does not contain $Z(D_{16})$.
8. (a) For each $m \geq 2$, find a group of length m .
 - (b) For each $m \geq 2$, find a group of length 1 with a subgroup of length m .
9. Let $G = K_0 \times K_1 \times \cdots \times K_r$, where each K_i is simple. Show that the groups K_i are the composition factors of G .
10. Describe the groups of length 2 by using Exercise 6 §8.1.
11. Let $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ be any subnormal series. If each factor G_i/G_{i+1} is finite, show that G is finite and that $|G| = |G_0/G_1| \cdot |G_1/G_2| \cdots |G_{n-1}/G_n|$.
12. If p is a prime, show that a finite group is a p -group if and only if all its composition factors are isomorphic to C_p .
13. Suppose that G is a group with a composition series. Show that any subnormal series $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ can be refined (by inserting groups if necessary) to a composition series for G .
14. Suppose that G has a composition series with no two factors isomorphic.
 - (a) Show that no two proper normal subgroups of G are isomorphic. [Hint: If $H \triangleleft G$ and $K \triangleleft G$, find a composition series through HK , H , and $H \cap K$ and one through HK , K , and $H \cap K$. Use Exercise 13.]
 - (b) Show that every normal subgroup of G is characteristic in G .
15. Let $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, where the p_i are distinct primes and each $n_i \geq 1$.
 - (a) Show that C_n has exactly r maximal normal subgroups.
 - (b) If $m = n_1 + n_2 + \cdots + n_r$, show that C_n has $\frac{m!}{n_1!n_2!\cdots n_r!}$ composition series. [Hint: induct on m .]
16. Prove the **Zassenhaus lemma**:¹⁰⁰ Let $H_1 \triangleleft H$ and $K_1 \triangleleft K$ be subgroups of a group G . Then $H_1(H \cap K_1) \triangleleft H_1(H \cap K)$, $K_1(H_1 \cap K) \triangleleft K_1(H \cap K)$, and we have $\frac{H_1(H \cap K)}{H_1(H \cap K_1)} \cong \frac{K_1(H \cap K)}{K_1(H_1 \cap K)}$. [Hint: Each group is isomorphic to $\frac{H \cap K}{(H_1 \cap K)(H \cap K_1)}$ by Theorem 4 §2.10.]
17. Prove the **Schreier refinement theorem**:¹⁰¹ Two subnormal series $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ and $H = H_0 \supseteq H_1 \supseteq \cdots \supseteq H_m = 1$ can be refined (by inserting groups) in such a way that the resulting series are equivalent. [Hint: $G_i = G_{i+1}(G_i \cap H_0) \supseteq G_{i+1}(G_i \cap H_1) \supseteq \cdots \supseteq G_{i+1}(G_i \cap H_m) = G_{i+1}$. Do a similar construction with the H_j and use the Zassenhaus lemma.]
18. Use the Schreier refinement theorem to prove the Jordan–Hölder theorem.

¹⁰⁰Due to Hans Zassenhaus.

¹⁰¹Due to Otto Schreier.

9.2 SOLVABLE GROUPS

In Section 9.1, we were concerned with groups that admit a composition series, that is, a subnormal series in which all the factors are simple. Although those groups are of interest, we obtain an even more important class of groups when the factors are required to be abelian rather than simple.

A group G is called a **solvable group**¹⁰² if there exists a subnormal series

$$G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$$

such that each factor G_i/G_{i+1} is abelian. Such a series is called a **solvable series** for G . Note that $\{1\}$ is solvable.

Example 1. Every abelian group G is solvable because $G \supseteq \{1\}$ is a solvable series.

Example 2. If p is a prime, every finite p -group is solvable. In fact, it has a composition series in which each factor is isomorphic to C_p (Theorem 8 §8.2).

Example 3. D_n is solvable for each n . Indeed, D_n has a cyclic subgroup H of index 2, so $H \triangleleft D_n$ and $D_n \supseteq H \supseteq \{1\}$ is a solvable series.

Example 4. S_4 is solvable because $S_4 \supseteq A_4 \supseteq K \supseteq \{\varepsilon\}$ is a solvable series, where $K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$.

Example 5. If p and q are primes, show that any group of order pq is solvable.

Solution. Because G is not simple by the Sylow theorems (Example 5 §8.4), let $K \neq \{1\}$ be a proper normal subgroup. Then $|K| = p$ or $|K| = q$; either way $G \supset K \supset \{1\}$ is a solvable series with factors C_p and C_q . \square

Suppose that G_1 is an abelian group. It is difficult to describe how to construct all groups G_2 such that $G_1 \triangleleft G_2$ and G_2/G_1 is abelian. Nonetheless, suppose we carry out this construction and then repeat it to construct a group G_3 such that $G_2 \triangleleft G_3$ and G_3/G_2 is abelian. If we continue this procedure, each group constructed is clearly solvable, and we can obtain every solvable group in this way—constructed from the bottom up, as it were. Viewing solvable groups in this way is useful, but an analogous top-down construction is actually more important. It is based on the derived subgroup introduced in Section 2.9.

Recall that an element of the form $aba^{-1}b^{-1}$ in a group G is called a commutator, denoted $[a, b]$. The set G' of all products of commutators is a subgroup, called the *derived subgroup* of G , and has the following properties (Theorem 3 §2.9):

- (1) $G' \triangleleft G$ and G/G' is abelian.
- (2) If $K \triangleleft G$ and G/K is abelian, then $G' \subseteq K$.

For a group G , repeatedly taking the derived subgroup leads to a subnormal series of subgroups $G \supseteq G' \supseteq G'' \supseteq G''' \supseteq \cdots$ in which each factor is abelian. There is a standard notation for these subgroups. Given a group G , construct subgroups $G^{(0)}$, $G^{(1)}$, $G^{(2)}$, \dots of G as follows:

- (1) Define $G^{(0)} = G$.
- (2) If $G^{(i)}$ has been constructed for $i \geq 0$, define $G^{(i+1)} = [G^{(i)}]'$.

¹⁰²These groups are also called *soluble* groups.

Thus, $G^{(1)} = G'$, $G^{(2)} = G''$, $G^{(3)} = G'''$, and so on. Furthermore, $G^{(i+1)} \triangleleft G^{(i)}$ for each i and the subnormal series

$$G = G^{(0)} \supseteq G^{(1)} \supseteq G^{(2)} \supseteq \dots$$

is called the **derived series** for G . Note that $G^{(i)}/G^{(i+1)}$ is abelian for each i by Theorem 3 §2.9. The groups $G^{(i)}$ are called the **higher derived subgroups** of G , and they are actually normal in G as the next theorem shows. The proof requires the following lemma; we leave the proof as Exercise 8.

Lemma 1. *Let G denote a group and let H be a subgroup.*

- (1) *If $\alpha : G \rightarrow G$ is a homomorphism, then $\alpha(G') \subseteq G'$.*
- (2) *$G' \subseteq H$ if and only if $H \triangleleft G$ and G/H is abelian.*
- (3) *If $H \subseteq G$ is a subgroup, then $H' \subseteq G'$.*

Theorem 1. *If G is a group, we have $G^{(i)} \triangleleft G$ for all $i \geq 0$.*

Proof. We use induction on $i \geq 0$. It is clear if $i = 0$, so assume inductively that $G^{(i)} \triangleleft G$ for some i . If $a \in G$ then $\sigma_a(G^{(i)}) = G^{(i)}$, where σ_a is the inner automorphism of G determined by a . But then Lemma 1 gives

$$\sigma_a[G^{(i+1)}] = \sigma_a[(G^{(i)})'] \subseteq (G^{(i)})' = G^{(i+1)}.$$

This shows that $G^{(i+1)} \triangleleft G$ and so completes the induction. ■

The solvable groups G are just those for which the derived series reaches $\{1\}$.

Theorem 2. *A group G is solvable if and only if $G^{(n)} = \{1\}$ for some $n \geq 1$.*

Proof. If $G^{(n)} = \{1\}$, then

$$G = G^{(0)} \supseteq G^{(1)} \supseteq G^{(2)} \supseteq \dots \supseteq G^{(n)} = \{1\}$$

is a solvable series for G because $G^{(i)}/G^{(i+1)} = G^{(i)}/[G^{(i)}]'$ is abelian for each i . Conversely, let $G = G_0 \supseteq G_1 \supseteq \dots \supseteq G_n = \{1\}$ be a solvable series for G . It suffices to show that $G^{(i)} \subseteq G_i$ holds for each i . This is clear if $i = 0$, so assume that $G^{(i)} \subseteq G_i$ for some $i \geq 0$. As G_i/G_{i+1} is abelian, we have $G'_i \subseteq G_{i+1}$. Hence,

$$G^{(i+1)} = [G^{(i)}]' \subseteq G'_i \subseteq G_{i+1}$$

by Lemma 1, which completes the induction. ■

If G is solvable, then $G^{(i)} \supsetneq G^{(i+1)}$ is strictly for all i by Corollary 1 of Theorem 3.

Theorem 2 provides a quick method of establishing several basic properties of solvable groups. We begin with the following result.

Theorem 3. *Every subgroup and image of a solvable group is again solvable.*

Proof. Suppose that G is solvable and let $G^{(n)} = \{1\}$. If H is a subgroup of G , it suffices to show that $H^{(i)} \subseteq G^{(i)}$ for each i . This follows by induction: It is clear when $i = 0$, and if $H^{(i)} \subseteq G^{(i)}$, then Lemma 1 gives

$$H^{(i+1)} = [H^{(i)}]' \subseteq [G^{(i)}]' = G^{(i+1)}.$$

Now let $\alpha : G \rightarrow K$ be an onto group homomorphism. It suffices to show that $K^{(i)} \subseteq \alpha[G^{(i)}]$ for each i . This is clear if $i = 0$ because α is onto, so assume that

$K^{(i)} \subseteq \alpha(G^{(i)})$. Then, given x and y in $K^{(i)}$, write $x = \alpha(a)$ and $y = \alpha(b)$, where $a, b \in G^{(i)}$. Hence,

$$[x, y] = [\alpha(a), \alpha(b)] = \alpha([a, b]) \in \alpha(G^{(i+1)}),$$

so $K^{(i+1)} \subseteq \alpha(G^{(i+1)})$ because $K^{(i+1)} = (K^{(i)})'$. \blacksquare

Corollary 1. If $H \neq \{1\}$ is a subgroup of a solvable group G , then $H' \neq H$.

Proof. If $H' = H$, then $H^{(2)} = [H']' = H' = H$ and an induction shows that $H^{(i)} = H \neq \{1\}$ holds for each i . As H is solvable by Theorem 3, this result contradicts Theorem 2. \blacksquare

Corollary 2. A simple group is solvable if and only if it is abelian (of prime order).

Proof. Let G be solvable. If G is simple, then either $G' = \{1\}$ (so G is abelian) or $G' = G$, (contradicting Corollary 1). The converse is clear. \blacksquare

Example 6. If $n \geq 5$, the symmetric group S_n is not solvable. For if S_n were solvable, A_n would be solvable by Theorem 3 so, as A_n is simple, Corollary 2 would imply that A_n is abelian, which is not the case. Hence, S_n is not solvable.

Example 6 explains the origin of the term *solvable*. A classical problem in the theory of equations was to find a formula for the roots of a real polynomial $x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ in terms of the coefficients a_i . If $n = 2$, the solution is the famous quadratic formula: $\frac{1}{2} \left[-a_1 \pm \sqrt{a_1^2 - 4a_0} \right]$. In general, such a formula should give the roots in terms of the coefficients a_i using only arithmetic operations and the extraction of roots. Such formulas were found for $n = 3$ and $n = 4$, but the case $n = 5$ proved to be difficult. Call a polynomial f solvable if such a formula exists. It will be shown in Chapter 10 that f is solvable if and only if a certain group (called the Galois group of f) is a solvable group. For example, the polynomial $x^5 - 6x + 2$ has Galois group S_5 (Example 1 §10.3) and so cannot be solvable. Incidentally, the first proof that a nonsolvable polynomial exists was given in 1824 by the young Norwegian mathematician Niels Henrik Abel, building on the work of Paolo Ruffini.

Theorem 4 gives a useful way to show that a group is solvable.

Theorem 4. If $K \triangleleft G$, then G is solvable if and only if both K and G/K are solvable.

Proof. Assume that K and G/K are solvable and let

$$K = K_0 \supseteq K_1 \supseteq \cdots \supseteq K_m = \{1\} \quad \text{and} \quad \frac{G}{K} = \frac{G_0}{K} \supseteq \frac{G_1}{K} \supseteq \cdots \supseteq \frac{G_r}{K} = \{K\}$$

be solvable series. Then

$$G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_r = K = K_0 \supseteq K_1 \supseteq \cdots \supseteq K_m = \{1\}$$

is a subnormal series for G and the factors are abelian because $\frac{G_i}{G_{i+1}} \cong \frac{G_i/K}{G_{i+1}/K}$ for each i . Hence, G is solvable. The converse follows by Theorem 3. \blacksquare

Example 7. If G_1, G_2, \dots, G_n are groups, then $G_1 \times G_2 \times \cdots \times G_n$ is solvable if and only if the same is true of each G_i .

Solution. By induction, it suffices to prove it for $n = 2$. Let $\theta : G_1 \times G_2 \rightarrow G_2$ be the projection given by $\theta(g_1, g_2) = g_2$. Then $(G_1 \times G_2)/\ker \theta \cong \theta(G_1 \times G_2) = G_2$ and $\ker \theta = G_1 \times \{1\} \cong G_1$ are both solvable, so Theorem 4 applies. \square

The above theorems are valid for arbitrary groups. We now give some conditions equivalent to solvability in a finite group.

Theorem 5. *The following conditions are equivalent for a finite group G .*

- (1) G is solvable.
- (2) The composition factors of G are all abelian.
- (3) $H' \neq H$ for every subgroup $H \neq \{1\}$ of G .

Proof. Note first that G has a composition series because it is finite.

(1) \Rightarrow (2). Each composition factor is simple and solvable (Theorem 3), and so is abelian by Corollary 2 of Theorem 3.

(2) \Rightarrow (3). Any composition series for G is a solvable series by (2).

(3) \Rightarrow (1). The derived series $G = G^{(0)} \supseteq G^{(1)} \supseteq G^{(2)} \supseteq \dots$ reaches $\{1\}$ because G is finite and $G^{(i)} \supset (G^{(i)})' = G^{(i+1)}$ for each i by (3). \blacksquare

Example 8. Let R be any ring. If $n \geq 3$, show that the group G of all invertible $n \times n$ matrices over R is not solvable.

Solution. Let E_{ij} denote the $n \times n$ matrix with (i, j) -entry 1 and zeros elsewhere. Then $E_{ij}E_{jk} = E_{ik}$, whereas $E_{ij}E_{lk} = 0$ if $j \neq l$. If I is the $n \times n$ identity matrix, this shows that $I + E_{ij}$ is in G whenever $i \neq j$ and that $(I + E_{ij})^{-1} = I - E_{ij}$. Now let H be the subgroup of G generated by the matrices $I + E_{ij}$, $i \neq j$. If i , j , and k are distinct indices (they exist because $n \geq 3$), compute

$$\begin{aligned} & (I + E_{ik})(I + E_{kj})(I + E_{ik})^{-1}(I + E_{kj})^{-1} \\ &= (I + E_{ik} + E_{kj} + E_{ij})(I - E_{ik} - E_{kj} + E_{ij}) \\ &= I + E_{ij}. \end{aligned}$$

This shows that every generator of H is a commutator from H and hence $H' = H$. Thus, G is not solvable by Corollary 1 of Theorem 3. \square

If F is a field, Example 8 shows that the general linear group $GL_n(F)$ of all $n \times n$ invertible matrices over F is not solvable if $n \geq 3$. If F is finite, Theorem 5 shows that a nonabelian simple group is lurking among the composition factors of $GL_n(F)$. In fact, such a group exists even if F is infinite. The mapping $A \mapsto \det A$ is an onto homomorphism $GL_n(F) \rightarrow F^*$ and the kernel is the special linear group $SL_n(F)$ of all matrices with determinant 1. It is not difficult to verify that the center of $SL_n(F)$ consists of all scalar matrices aI , where $a \in F$ satisfies $a^n = 1$. The factor group

$$PSL_n(F) = \frac{SL_n(F)}{Z[SL_n(F)]}$$

is called the **projective special linear group** (of degree n) over F . These groups comprise another infinite family of finite, simple, nonabelian groups (in addition to the alternating groups A_n , $n \geq 5$). The theorem was proved in 1870 by Camille Jordan for \mathbb{Z}_p , p a prime and, in early 1900, Leonard Eugene Dickson proved it for all finite fields.

Theorem 6. Jordan–Dickson Theorem. If F is a finite field, then $PSL_n(F)$ is a finite nonabelian simple group for all $n \geq 2$, except for $PSL_2(\mathbb{Z}_2)$ and $PSL_2(\mathbb{Z}_3)$.

The proof is beyond the scope of this book.¹⁰³

The class of solvable groups is large. Of course, it contains all abelian groups, and a celebrated theorem of William Burnside asserts that every group of order $p^n q^m$ is solvable, where p and q are primes. In a different direction, Georg Frobenius showed that every group of square-free order is solvable. In 1911, Burnside conjectured that every nonabelian finite simple group has even order, equivalently (Exercise 13) that every group of odd order is solvable. This conjecture remained an open question until 1963 when two American algebraists Walter Feit and John Thompson proved that it is true. The proof is 254 pages long and fills an entire issue of the *Pacific Journal of Mathematics*,¹⁰⁴ and it is widely regarded as the best single paper in finite group theory. Thompson went on to classify all minimal finite simple groups, that is, those in which every proper subgroup is solvable, and played an important role in the classification of all finite simple groups. He was awarded the Fields Medal in 1970, the highest honor a mathematician can attain.

Even though the class of solvable groups is very large, many theorems are true for solvable groups that are not true of groups in general. One such theorem, a fundamental strengthening of the Sylow theorems in any solvable group, was first proved in 1928 by the British mathematician Philip Hall.

Theorem 7. Hall's Theorem. Let G be a group of order nm , where n and m are relatively prime. If G is solvable, then

- (1) G has a subgroup of order n and any two are conjugate.
- (2) If $k|n$, each subgroup of order k is contained in a subgroup of order n .

We omit the proof.¹⁰⁵ Hall went on to develop the theory of finite solvable groups and influenced an entire generation of group theorists.

Exercises 9.2

1. Is $Z(G) \neq \{1\}$ for every solvable group $G \neq \{1\}$? Support your answer.
2. If G is solvable, is $N(H) \neq H$ for each subgroup $H \neq G$? Support your answer.
3. Is G' abelian for every solvable group G ? Support your answer.
4. Does every solvable group of order n have a subgroup of order m for each divisor m of n ? Support your answer.
5. Give an example of a nonsolvable group in which every Sylow subgroup is abelian.
6. Show that a nonsolvable group of minimal order must be simple.

¹⁰³See Kargapolov, M.I. and Merzljakov, J.I., *Fundamentals of the Theory of Groups*, Springer-Verlag, 1979; Rotman, J.J., *The Theory of Groups: An Introduction*, 2nd ed., Boston: Allyn & Bacon, 1973; Artin, E., *Geometric Algebra*, New York: Interscience, 1957. For an elementary proof when $n = 2$, see Lang, S., *Undergraduate Algebra*, Berlin: Springer-Verlag, 1987.

¹⁰⁴Feit, W. and Thomson, J.G., *Solvability of groups of odd order*, *Pacific Journal of Mathematics*, 13 (1963), 775–1029.

¹⁰⁵See Kargapolov, M.I. and Merzljakov, J.I., *Fundamentals of the Theory of Groups*, Springer-Verlag, 1979; MacDonald, I.D., *The Theory of Groups*, London: Oxford University Press, 1968; Rotman, J.J., *The Theory of Groups: An Introduction*, 2nd ed., Boston: Allyn & Bacon, 1973.

7. Suppose G has a solvable, maximal normal subgroup. Either prove G is solvable or give a counterexample.
8. Prove Lemma 1.
9. If $|G| = p^2q$, p, q primes, show that G is solvable. [Hint: Exercise 14 §8.4.]
10. If $|G| = p^2q^2$, p, q primes, show that G is solvable. [Hint: Example 9 §8.4 and the preceding exercise.]
11. (a) Show that

$$G = \left\{ \begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} \mid a, b, c \in F \right\}$$

is a solvable group for any field F .

(b) Show that

$$G = \left\{ \begin{bmatrix} x & a & b \\ 0 & y & c \\ 0 & 0 & z \end{bmatrix} \mid x, y, z, a, b, c \in F; xyz \neq 0 \right\}$$

is a solvable group for any field F .

12. If p and q are primes, show that every group of order $p^m q^n$ is solvable if and only if the only simple groups of this type are the cyclic groups of order p or q . (Burnside proved that these statements are true.)
13. Show that every group of odd order is solvable if and only if every finite nonabelian simple group has even order. (These statements are true by Feit and Thompson.)
14. Find the composition length of a solvable group of order $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, where p_i are distinct primes. [Hint: Example 8 §9.1.]
15. Show that a solvable group is finite if and only if it has a composition series.
16. Show that the following are equivalent for a group G .
 - (a) G is solvable.
 - (b) G' is solvable.
 - (c) $G/Z(G)$ is solvable.
17. If $H \triangleleft G$ and $K \triangleleft G$ show that $G/(H \cap K)$ is solvable if and only if the product $(G/H) \times (G/K)$ is solvable.
18. If $K_i \triangleleft G$ for $i = 1, 2, \dots, n$, put $K = K_1 \cap K_2 \cap \cdots \cap K_n$. If G/K_i is solvable for each i , show that G/K is solvable.
19. If H and K are solvable subgroups of G and $K \triangleleft G$, show that HK is solvable.
20. (a) If G is a finite group and $Z(G/K)$ is nontrivial for all $K \triangleleft G$, $K \neq G$, show that G is solvable.
 (b) Show that the converse of (a) is false for a finite group G .
21. If $G \neq 1$ is solvable, show that
 - (a) G has a nontrivial abelian factor group.
 - (b) G has a nontrivial abelian normal subgroup.
22. Show that the following are equivalent for a nontrivial finite group G .
 - (1) G is solvable.
 - (2) Every nontrivial normal subgroup of G has a nontrivial abelian factor group.
 - (3) Every nontrivial factor group of G has a nontrivial abelian normal subgroup.
23. If G is a finite group, define $R = R(G) = \bigcap \{K \triangleleft G \mid G/K \text{ is solvable}\}$.
 - (a) Show that $R = R(G)$ is the smallest normal subgroup of G such that G/R is solvable. [Hint: Exercise 18.]
 - (b) Show that G is solvable if and only if $R(G) = \{1\}$.

- (c) If $\alpha : G \rightarrow H$ is a group homomorphism, show that $\alpha[R(G)] \subseteq R(H)$.
Hint: Consider $\{k \mid \alpha(k) \in R(H)\}$.
- (d) If $H \subseteq G$ is a subgroup, show that $R(H) \subseteq H \cap R(G)$, with equality if $H \triangleleft G$.
24. If G is a finite group, define $S(G) = \prod\{K \triangleleft G \mid K \text{ is solvable}\}$.
- (a) Prove $S(G)$ is the largest solvable, normal subgroup of G . [*Hint:* Exercise 19.]
 - (b) Prove G is solvable if and only if $S(G) = G$.
 - (c) If $\alpha : G \rightarrow H$ is an onto group homomorphism, show that $\alpha[S(G)] \subseteq S(H)$.
 - (d) If $H \subseteq G$ is a subgroup, show that $H \cap S(G) \subseteq S(H)$, with equality if $H \triangleleft G$.
 - (e) Show that $S(G/S) = \{1\}$.
25. A group G is called **polycyclic** if it has a solvable series with every factor cyclic.
- (a) Show that every finite solvable group is polycyclic.
 - (b) Show that every polycyclic group is finitely generated.
 - (c) Show that every subgroup and homomorphic image of a polycyclic group is polycyclic. [*Hint:* Lemma 1 §9.1.]
 - (d) If $K \triangleleft G$, show that G is polycyclic if and only if K and G/K are polycyclic.
 - (e) Show that the following are equivalent for a group G . [*Hint:* Theorem 3 §7.2.]
 - (i) G is polycyclic
 - (ii) Every subgroup of G is solvable and finitely generated.
 - (iii) Every normal subgroup of G is solvable and finitely generated.
26. A class \mathcal{V} of groups is called a **subvariety** if $\{1\} \in \mathcal{V}$ and each subgroup and homomorphic image of a group in \mathcal{V} is again in \mathcal{V} . Examples: abelian groups, p -groups for a fixed prime p and torsion groups (each element has finite order). If \mathcal{V} is a subvariety, a group G is called \mathcal{V} -solvable if there is a subnormal series $G = G_0 \supseteq G_1 \supseteq \dots \supseteq G_n = \{1\}$ with $G_i/G_{i+1} \in \mathcal{V}$ for each i . If G is a group and $K \triangleleft G$, show that G is \mathcal{V} -solvable if and only if both K and G/K are \mathcal{V} -solvable. [*Hint:* Lemma 1 §9.1.]
27. A subvariety \mathcal{V} of groups (Exercise 26) is called a **variety** if, in addition, $G \times H$ is in \mathcal{V} whenever G and H are in \mathcal{V} . Examples: abelian groups, p -groups for a fixed prime p , torsion groups, and \mathcal{V} -solvable groups, where \mathcal{V} is any subvariety (by Exercise 26). If \mathcal{V} is a variety and G is a finite group, the **\mathcal{V} -derived subgroup** of G is defined to be $\mathcal{V}(G) = \bigcap\{K \triangleleft G \mid G/K \text{ is in } \mathcal{V}\}$. Let G denote a finite group.
- (a) Show that $\mathcal{V}(G) \triangleleft G$ and $G/\mathcal{V}(G)$ is in \mathcal{V} .
 - (b) If $K \triangleleft G$, show that G/K is in \mathcal{V} if and only if $\mathcal{V}(G) \subseteq K$.
 - (c) If H is a subgroup of G , show that $\mathcal{V}(H) \subseteq \mathcal{V}(G)$. [*Hint:* Lemma 2 §9.1.]
 - (d) If $\alpha : G \rightarrow H$ is a homomorphism of groups, show that $\alpha[\mathcal{V}(G)] \subseteq \mathcal{V}(H)$.
28. If \mathcal{V} is a variety of finite groups, define $\mathcal{V}_0(G) = G$ and $\mathcal{V}_{k+1}(G) = \mathcal{V}[\mathcal{V}_k(G)]$ for each $k \geq 0$. Let G denote a finite group. Show that
- (a) If $\alpha : G \rightarrow H$ is a group homomorphism, then $\alpha[\mathcal{V}_k(G)] \subseteq \mathcal{V}_k(H)$ for all k .
 - (b) $\mathcal{V}_k(G) \triangleleft G$ for each k .
 - (c) G is \mathcal{V} -solvable if and only if $\mathcal{V}_k(G) = \{1\}$ for some k .
 - (d) Every subgroup of a \mathcal{V} -solvable group is \mathcal{V} -solvable.
 - (e) G is \mathcal{V} -solvable if and only if $\mathcal{V}(H) \neq H$ for all subgroups $H \neq \{1\}$ of G .

9.3 NILPOTENT GROUPS

If G is a group, the definition of the derived subgroup G' guarantees that G is abelian if and only if $G' = \{1\}$. If the process of taking the derived subgroup is

iterated, the derived series $G = G^{(0)} \supseteq G^{(1)} \supseteq G^{(2)} \supseteq \dots$ is obtained, and G is solvable if and only if this series (of normal subgroups of G) reaches $\{1\}$ in a finite number of steps (Theorem 2 §9.2). Note that $G^{(1)} = G'$. Now the center $Z(G)$ plays an analogous role to G' in the sense that G is abelian if and only if $Z(G) = G$. In view of this, an irresistible question arises: Is there a way to iterate the formation of the center so as to create a series $\{1\} = Z_0 \subseteq Z_1 \subseteq Z_2 \subseteq \dots$ of normal subgroups of G (with $Z(G) = Z_1$) such that G is solvable if and only if this series reaches G in a finite number of steps? The answer is yes *and* no. Yes, there is a natural way to define such a series. No, it does not characterize the solvable groups in this way. Rather, it characterizes a smaller class of groups called the nilpotent groups. In this section, we define these groups and show that the finite ones are precisely the finite groups that are isomorphic to the direct product of their Sylow subgroups.

Central Series

If G is a group, define a series $Z_0(G), Z_1(G), Z_2(G), \dots$ of normal subgroups of G inductively as follows:

- (1) Take $Z_0(G) = \{1\}$.
- (2) If $Z_i(G) \triangleleft G$ has been constructed, define $Z_{i+1}(G)$ the unique normal subgroup of G that contains $Z_i(G)$ and satisfies $Z\left(\frac{G}{Z_i(G)}\right) = \frac{Z_{i+1}(G)}{Z_i(G)}$.

Then $Z_{i+1}(G) \supseteq Z_i(G)$ and $Z_i(G) \triangleleft Z_{i+1}(G)$ for each $i \geq 0$, and the series

$$\{1\} = Z_0(G) \subseteq Z_1(G) \subseteq Z_2(G) \subseteq \dots$$

is called the **ascending central series** of G . Note that

$$Z_1(G) = Z(G) \quad \text{because} \quad Z\left(\frac{G}{\{1\}}\right) = \frac{Z_1(G)}{\{1\}}.$$

The ascending central series may never reach G , even if G is solvable:

Example 1. Suppose $Z(G) = \{1\}$ where $G \neq \{1\}$ —for example the solvable group S_3 . Then $Z_1(G) = Z(G) = \{1\}$, so we have $Z_2(G) \cong \frac{Z_2(G)}{Z_1(G)} = Z\left(\frac{G}{Z_1(G)}\right) \cong Z(G) = \{1\}$. Hence, $Z_2(G) = \{1\}$, and this process continues inductively to show that $Z_k(G) = \{1\}$ for all k .

To characterize the groups G for which the ascending central series reaches G , it is useful to define a related *descending* central series. This requires a new notion. Recall that the derived subgroup G' is generated by the commutators $[a, b] = aba^{-1}b^{-1}$ in G . We extend this idea as follows. If H and K are subgroups of a group G , define

$$[H, K] = \langle \{[h, k] \mid h \in H \text{ and } k \in K\} \rangle$$

to be the subgroup generated by the commutators $[h, k]$, with $h \in H$ and $k \in K$. Note that $[h, k]^{-1} = [k, h]$ for all $h \in H$ and $k \in K$. Hence, $[H, K] = [K, H]$, and this group consists of all products of commutators of the form $[h, k]$ or $[k, h]$, where $h \in H$ and $k \in K$. In particular, $[G, G] = G'$. Lemma 1 collects several other useful facts about these subgroups.

Lemma 1. Let H, K, H_1 , and K_1 be subgroups of a group G .

- (1) $[H, K] = [K, H]$.
- (2) If $H \subseteq H_1$ and $K \subseteq K_1$, then $[H, K] \subseteq [H_1, K_1]$.

- (3) If $H \triangleleft G$ and $K \triangleleft G$, then $[H, K] \triangleleft G$.
- (4) $H \triangleleft G$ if and only if $[H, G] \subseteq H$.
- (5) Suppose that $K \subseteq H \subseteq G$ and $K \triangleleft G$. Then

$$H/K \subseteq Z(G/K) \quad \text{if and only if} \quad [H, G] \subseteq K.$$

Proof. We prove (5) and leave the rest as Exercise 2. If $\frac{H}{K} \subseteq Z\left(\frac{G}{K}\right)$, then we have $hKhK = gKhK$ for all $g \in G$ and $h \in H$; that is, $[h, g] \in K$. Hence, $[H, G] \subseteq K$. Since this argument works in reverse, we have proved (5). ■

Now, given a group G , define a series $\Gamma_0(G), \Gamma_1(G), \Gamma_2(G), \dots$ of subgroups of G inductively as follows:

- (1) Take $\Gamma_0(G) = G$.
- (2) If $\Gamma_i(G)$ has been constructed, define $\Gamma_{i+1}(G) = [\Gamma_i(G), G]$.

Then $\Gamma_i(G) \triangleleft G$ for all $i \geq 0$ by induction on i (using (3) of Lemma 1), whence $\Gamma_{i+1}(G) = [\Gamma_i(G), G] \subseteq \Gamma_i(G)$ for each i by (4) of Lemma 1. Thus, we obtain a series of normal subgroups

$$G = \Gamma_0(G) \supseteq \Gamma_1(G) \supseteq \Gamma_2(G) \supseteq \dots$$

This is called the **descending central series** of G . The name comes from the fact that, using (5) of Lemma 1,

$$\frac{\Gamma_i(G)}{\Gamma_{i+1}(G)} \subseteq Z\left[\frac{G}{\Gamma_{i+1}(G)}\right]$$

holds for each $i \geq 0$. Note that

$$\Gamma_1(G) = [G, G] = G' \quad \text{is the derived subgroup of } G.$$

If G is an abelian group, then $Z_1(G) = G$ and $\Gamma_1(G) = \{1\}$. On the other hand, there are groups (even solvable ones by Example 1) for which the ascending central series does not reach G and the descending central series does not reach 1. However, if either possibility occurs, so does the other.

Lemma 2. *The following are equivalent for a group G and an integer $n \geq 0$:*

- (1) $\Gamma_n(G) = \{1\}$.
- (2) $Z_n(G) = G$.
- (3) A series $G = G_0 \supseteq G_1 \supseteq \dots \supseteq G_n = \{1\}$ exists with $G_i \triangleleft G$ for each i and $G_i/G_{i+1} \subseteq Z(G/G_{i+1})$.

Proof. Write $\Gamma_i(G) = \Gamma_i$ and $Z_i(G) = Z_i$ for each i .

(1) \Rightarrow (2). If $\Gamma_n = \{1\}$, we show that $\Gamma_{n-i} \subseteq Z_i$ for each $i = 0, 1, 2, \dots$ (so $Z_n = G$). This is clear if $i = 0$ by (1), so assume $\Gamma_{n-i} \subseteq Z_i$, where $i > 0$. If $a \in \Gamma_{n-i-1}$, then, for all $g \in G$, $[a, g] \in [\Gamma_{n-i-1}, G] = \Gamma_{n-i} \subseteq Z_i$. Thus, aZ_i is in the center of G/Z_i and so $a \in Z_{i+1}$. Hence, $\Gamma_{n-i-1} \subseteq Z_{i+1}$ as required.

(2) \Rightarrow (3). Given (2), use $G = Z_n \supseteq Z_{n-1} \supseteq \dots \supseteq Z_0 = \{1\}$ in (3).

(3) \Rightarrow (1). Given (3), we show that $\Gamma_i \subseteq G_i$ for each $i = 0, 1, 2, \dots$ (so $\Gamma_n = \{1\}$). This is clear if $i = 0$, so assume that $\Gamma_i \subseteq G_i$ for some $i > 0$. We must show that $[\Gamma_i, G] = \Gamma_{i+1} \subseteq G_{i+1}$, so we show that $[a, g] \in G_{i+1}$ for all $a \in \Gamma_i$, $g \in G$. But $\frac{G_i}{G_{i+1}} \subseteq Z\left(\frac{G}{G_{i+1}}\right)$ by (3), so $a \in \Gamma_i \subseteq G_i$ implies that aG_{i+1} commutes with gG_{i+1} for all $g \in G$. This in turn implies that $[a, g] \in G_{i+1}$, as required. ■

A group G is called a **nilpotent group** if the conditions in Lemma 2 are satisfied for some $n \geq 0$. The smallest integer n for which $\Gamma_n(G) = \{1\}$, equivalently $Z_n(G) = G$, is called the **nilpotency class** of G . Thus, if G is nilpotent, then G has class 1 if and only if it is abelian, and (Exercise 11) G has class 2 if and only if it is nonabelian and $G' \subseteq Z(G)$.

A series as in (3) of Lemma 2 is called a **central series** for G . Suppose that $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ is any central series for a nilpotent group G . Then the proof that (3) \Rightarrow (1) in Lemma 2 derives the first of the following inclusions:

$$\Gamma_i(G) \subseteq G_i \subseteq Z_{n-i}(G) \quad \text{for } 0 \leq i \leq n.$$

We leave the second inclusion for the reader (Exercise 7). Hence, we often call the series $G = \Gamma_0(G) \supseteq \Gamma_1(G) \supseteq \Gamma_2(G) \supseteq \cdots$ and $1 = Z_0(G) \subseteq Z_1(G) \subseteq Z_2(G) \subseteq \cdots$ the **lower** and **upper central series**, respectively.

Example 2. Every abelian group is nilpotent.

Example 3. If p is a prime, every finite p -group is nilpotent. In fact, if we write $Z_i(G) = Z_i$ for each i , Theorem 6 §8.2 shows that $Z_{i+1}/Z_i = Z(G/Z_i)$ is not trivial if $Z_i \neq G$ because G/Z_i is a p -group. Hence, $1 \subset Z_1 \subset Z_2 \subset \cdots$, which eventually reaches G because G is finite.

Example 4. Show that every nilpotent group G is solvable, but not conversely.

Solution. If G is nilpotent, the series $\{1\} = Z_0(G) \subseteq \cdots \subseteq Z_n(G) = G$ is a solvable series. However, S_3 is solvable but not nilpotent by Example 1. \square

Theorem 1. Every subgroup and image of a nilpotent group is again nilpotent.

Proof. Let G be nilpotent. To show that a subgroup $K \subseteq G$ is nilpotent, it suffices to show that $\Gamma_i(K) \subseteq \Gamma_i(G)$ for each i . This is clear if $i = 0$. If $\Gamma_i(K) \subseteq \Gamma_i(G)$ for some i , then the induction goes through because

$$\Gamma_{i+1}(K) = [\Gamma_i(K), K] \subseteq [\Gamma_i(G), G] = \Gamma_{i+1}(G).$$

Now let $\alpha : G \rightarrow H$ be an onto homomorphism; we show that $\Gamma_i(H) \subseteq \alpha[\Gamma_i(G)]$ for each i . If $i = 0$, it is because α is onto. In general, let $\Gamma_i(H) \subseteq \alpha[\Gamma_i(G)]$, and let $y \in \Gamma_i(H)$ and $h \in H$. Write $y = \alpha(x)$, where $x \in \Gamma_i(G)$, and (as α is onto) write $h = \alpha(g)$, $g \in G$. Then

$$[y, h] = [\alpha(x), \alpha(g)] = \alpha[x, g] \in \alpha[\Gamma_i(G), G] = \alpha(\Gamma_{i+1}(G)).$$

Hence, $\Gamma_{i+1}(H) = [\Gamma_i(H), H] \subseteq \alpha(\Gamma_{i+1}(G))$, as required. \blacksquare

Corollary. A group G is nilpotent if and only if G' is nilpotent.

Proof. Write $D = G'$. If $D^{(n)} = \{1\}$, then $G^{(n+1)} = D^{(n)} = \{1\}$. \blacksquare

By Theorem 1, if $K \triangleleft G$ and G is nilpotent, then both K and G/K are nilpotent. The converse is false (S_3 is again a counterexample) in contrast to the situation for solvable groups. However, the converse does hold when $K \subseteq Z(G)$.

Theorem 2. If G is a group, $K \subseteq Z(G)$, and G/K is nilpotent, then G is nilpotent.

Proof. Assume that G/K is nilpotent, and let $\theta : G \rightarrow G/K$ be the coset map. By hypothesis, let $\theta(G) = X_0 \supseteq X_1 \supseteq \cdots \supseteq X_n = \{K\}$ be a central series for $\theta(G)$,

where $\frac{X_i}{X_{i+1}} \subseteq Z(\frac{\theta(G)}{X_{i+1}})$ for $0 \leq i \leq n - 1$. Write $X_i = G_i/K = \theta(G_i)$ for $0 \leq i \leq n$. Then we obtain the series

$$G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = K \supseteq G_{n+1} = \{1\}.$$

We show this is a central series for G . First, $\frac{G_n}{G_{n+1}} = \frac{K}{\{1\}} \subseteq Z(\frac{G}{\{1\}}) = \frac{Z(G)}{\{1\}}$ because $K \subseteq Z(G)$. To see that $\frac{G_i}{G_{i+1}} \subseteq Z(\frac{G}{G_{i+1}})$ for $0 \leq i < n$, let $a \in G_i$. Then $\theta(a) \in \theta(G_i) = X_i$, so $\theta(a)X_{i+1}$ commutes with $\theta(g)X_{i+1}$ for all $g \in G$. Thus, $\theta[a, g] = [\theta(a), \theta(g)] \in X_{i+1} = \theta(G_{i+1})$, say $\theta[a, g] = \theta(b)$, $b \in G_{i+1}$. Thus, $[a, g]b^{-1} \in \ker \theta = K \subseteq G_{i+1}$, so $[a, g] \in G_{i+1}b = G_{i+1}$. This means aG_{i+1} commutes with gG_{i+1} , that is, $aG_{i+1} \in Z(\frac{G}{G_{i+1}})$, as required. ■

The next result will be needed in Theorem 4.

Theorem 3. *If G_1, G_2, \dots, G_n are nilpotent, so also is $G_1 \times G_2 \times \cdots \times G_n$.*

Proof. This follows because $\Gamma_i(G_1 \times G_2 \times \cdots \times G_n) \subseteq \Gamma_i(G_1) \times \cdots \times \Gamma_i(G_n)$ for each i , a fact that we leave as Exercise 6. ■

Theorem 3 and Example 3 combine to show that any finite direct product of finite p -groups (for various primes p) is nilpotent. In fact, every finite nilpotent group is isomorphic to such a direct product. We need the following notion.

A subgroup M of a group G is said to be **maximal** in G if $M \neq G$ and the only subgroups H such that $M \subseteq H \subseteq G$ are $H = M$ and $H = G$. Clearly, every proper subgroup K of a finite group is contained in a maximal subgroup—one of maximal order containing K . If G is finite, every subgroup of prime index is maximal by Example 6 §2.6.¹⁰⁶ The converse is not necessarily true (any subgroup of index 4 in A_4 is maximal), but it does hold in a finite p -group. Moreover, in this case, the maximal subgroups (of index p) are necessarily normal (see the corollary to Theorem 1 §8.3). This property characterizes the finite nilpotent groups.

Theorem 4. Burnside–Wielandt Theorem.¹⁰⁷ *The following conditions are equivalent for a finite group $G \neq \{1\}$:*

- (1) G is nilpotent.
- (2) $N(H) \neq H$ for all subgroups $H \neq G$ of G .
- (3) Every maximal subgroup of G is normal in G .
- (4) Every Sylow subgroup of G is normal in G .
- (5) G is isomorphic to the direct product of its Sylow subgroups.

Proof. (1) \Rightarrow (2). Write $Z_i = Z_i(G)$ for each i and assume that $Z_n = G$. If $H \neq G$ is a subgroup of G , then $Z_0 \subseteq H$ but $Z_n \not\subseteq H$, so an integer $k \geq 0$ exists such that $Z_k \subseteq H$ but $Z_{k+1} \not\subseteq H$. Choose $a \in Z_{k+1}$, $a \notin H$. Then aZ_k is in the center of G/Z_k , so if $h \in H$, aZ_k and hZ_k commute. Hence, $hah^{-1}a^{-1} \in Z_k \subseteq H$, from which $aHa^{-1} \subseteq H$. Thus, $a \in N(H)$, and so $N(H) \neq H$.

(2) \Rightarrow (3). Let M be a maximal subgroup of G . Since $M \subseteq N(M) \subseteq G$, (2) implies that $N(M) = G$. Hence, $M \triangleleft G$.

¹⁰⁶This is also true if G is infinite by Exercise 31, §2.6.

¹⁰⁷The name honors William Burnside and Helmut Wielandt.

(3) \Rightarrow (4). Suppose P is a nonnormal Sylow p -subgroup of G . Then $N(P) \neq G$, so let $N(P) \subseteq M$, where M is a maximal subgroup of G . Because $P \subseteq M$, (3) gives $aPa^{-1} \subseteq aMa^{-1} = M$ for all $a \in G$. Hence, both P and aPa^{-1} are Sylow p -subgroups of M and so are conjugate in M , say $P = m(aPa^{-1})m^{-1}$ for some $m \in M$. But then $ma \in N(P)$, so $a \in M$. Because $a \in G$ was arbitrary, this means $G \subseteq M$, a contradiction. This proves (4).

(4) \Rightarrow (5). Let P_1, P_2, \dots, P_r denote the distinct Sylow subgroups of G .

Claim 1. $P_1P_2 \cdots P_k \cong P_1 \times P_2 \times \cdots \times P_k$ for each $k = 2, 3, \dots, r$.

Proof. It is clear if $k = 1$. Assume inductively that $P_1P_2 \cdots P_k \cong P_1 \times P_2 \times \cdots \times P_k$ for some $k > 1$. Then $(P_1P_2 \cdots P_k) \cap P_{k+1} = \{1\}$ because elements in the two subgroups have relatively prime orders. By Theorem 6 §2.8,

$$(P_1P_2 \cdots P_k)P_{k+1} \cong (P_1 \times P_2 \times \cdots \times P_k) \times P_{k+1} \cong P_1 \times P_2 \times \cdots \times P_{k+1}$$

because $P_1P_2 \cdots P_k \triangleleft G$ and $P_{k+1} \triangleleft G$. This proves the claim.

The claim gives $|P_1P_2 \cdots P_r| \cong |P_1||P_2| \cdots |P_r| = |G|$. Hence $G = P_1P_2 \cdots P_r$ and (5) follows, again by the Claim.

(5) \Rightarrow (1). This follows from Theorem 3 and Example 3. ■

It is interesting to compare (2) in Theorem 4 with the result (Theorem 5 §9.2) that a finite group G is solvable if and only if $H' \neq H$ for every subgroup $H \neq \{1\}$.

Since every finite abelian group is nilpotent, the implication (1) \Rightarrow (5) in Theorem 4 gives another proof of the primary decomposition theorem for finite abelian groups (Corollary 2 of Theorem 3 §7.2). We reformulate (1) \Leftrightarrow (5) as

Corollary 1. A finite group G is nilpotent if and only if G is isomorphic to a finite direct product of p -groups for various primes p .

Frattini and Fitting Subgroups

One of the most important aspects of the study of nilpotent groups is that every finite group G contains a nilpotent subgroup Φ , which is characteristic in G (that is, $\sigma(\Phi) = \Phi$ for every automorphism σ of G)—these subgroups are discussed in Corollary 3 of Theorem 3 §2.8). We now turn to a discussion of this.

If $G \neq \{1\}$ is a finite group, define the **Frattini subgroup**

$$\Phi(G) = \bigcap \{M \subseteq G \mid M \text{ is a maximal subgroup of } G\}.$$

Define $\Phi\{1\} = \{1\}$. This was introduced in 1885 by Giovanni Frattini.

Example 5. $\Phi(A_4) = \{\varepsilon\}$. Indeed, $K = \{\varepsilon, (1 \ 2)(3 \ 4), (1 \ 3)(2 \ 4), (1 \ 4)(2 \ 3)\}$ is maximal, being of index 3, and $M = \{\varepsilon, (1 \ 2 \ 3), (1 \ 3 \ 2)\}$ is maximal (it has index 4, but A_4 has no subgroup of index 2). Hence, we have $\Phi(A_4) \subseteq K \cap M = \{\varepsilon\}$.

Example 6. If $Q = \{\pm 1, \pm i, \pm j, \pm k\}$ is the quaternion group, then $\Phi(Q) = \{1, -1\}$ because $\langle i \rangle, \langle j \rangle, \text{ and } \langle k \rangle$ are the only maximal subgroups.

Example 7. If $G = \langle a \rangle$ and $o(a) = p^n$, where p is a prime, $\Phi(G) = \langle a^p \rangle$ because $\langle a^p \rangle$ is the unique maximal subgroup (of index p).

Theorem 5. Let G be a group and write $\Phi = \Phi(G)$. Then the following hold:

(1) If $\alpha : G \rightarrow H$ is an onto group homomorphism, then $\alpha(\Phi) \subseteq \Phi(H)$.

- (2) In particular, Φ is a characteristic subgroup of G .
(3) G is nilpotent if and only if $G' \subseteq \Phi$.

Proof. (1) If $U \subseteq H$ is a maximal subgroup, we must show that $\alpha(\Phi) \subseteq U$. If we define $M = \{m \in G \mid \alpha(m) \in U\}$, then it suffices to show that M is a maximal subgroup of G (since then $\Phi \subseteq M$). But if $M \subseteq K \subseteq G$ are subgroups of G then $\alpha(M) \subseteq \alpha(K) \subseteq \alpha(G)$. Since α is onto, this is $U \subseteq \alpha(K) \subseteq H$, so $\alpha(K) = U$ or $\sigma(K) = H$. These imply that $K = M$ or $K = G$.

(2) This follows from (1) if $H = G$ and α is an automorphism of G .

(3) $G' \subseteq \Phi$ if and only if $G' \subseteq M$ for each maximal subgroup M of G , if and only if $M \triangleleft G$ for each M , and if and only if G is nilpotent by Theorem 4. \blacksquare

Corollary 1. The following are equivalent for a finite group G :

- (1) G is nilpotent.
- (2) $G/\Phi(G)$ is abelian.
- (3) $G/\Phi(G)$ is nilpotent.

Proof. (1) \Leftrightarrow (2) restates (3) of Theorem 5 and (2) \Rightarrow (3) is obvious.

(3) \Rightarrow (1). Given (3), we show that every maximal subgroup M of G is normal. Write $\Phi = \Phi(G)$. Since $\Phi \subseteq M$, the subgroup M/Φ is maximal in G/Φ by the correspondence theorem, so $M/\Phi \triangleleft G/\Phi$ by (3) and Theorem 4. But then $M \triangleleft G$, again by the correspondence theorem. \blacksquare

To see that $\Phi(G)$ is a nilpotent group, we first characterize it in terms of the following concept. An element $t \in G$ is called a **nongenerator** in G if it can be omitted from any generating set X of G ; that is, if $G = \langle X \cup \{t\} \rangle$, then $G = \langle X \rangle$.

Theorem 6. Let G denote any finite group. Then the following hold:

- (1) $\Phi(G) = \{t \mid t \text{ is a nongenerator of } G\}$.
- (2) $\Phi(G)$ is a nilpotent group.

Proof. For convenience, write $\Phi = \Phi(G)$.

(1) Write $N = \{t \mid t \text{ is a nongenerator of } G\}$ and let $a \in \Phi$. If $a \notin N$, then $X \subseteq G$ exists such that $\langle X \cup \{a\} \rangle = G$ but $\langle X \rangle \neq G$. So let $\langle X \rangle \subseteq M$, where M is a maximal subgroup of G . Then $a \in M$ because $\Phi \subseteq M$, whence $G = \langle X \cup \{a\} \rangle \subseteq M$, a contradiction. Hence, $\Phi \subseteq N$. Conversely, if $t \in N$ and M is a maximal subgroup of G , then $g \in M$ (otherwise $\langle M \cup \{t\} \rangle = G$). Hence, $N \subseteq \Phi$.

(2) By Theorem 4 we show that every Sylow p -subgroup P of Φ is normal in G . If $g \in G$, then $gPg^{-1} \subseteq g\Phi g^{-1} = \Phi$ by (2) of Theorem 5. Hence, both gPg^{-1} and P are Sylow p -subgroups of Φ and so are conjugate in Φ , say $t(gPg^{-1})t^{-1} = P$, where $t \in \Phi$. Thus, $tg \in N(P)$, which yields $G = \Phi N(P)$. But then $G = \langle \Phi \cup N(P) \rangle$ so, as Φ is finite, $G = \langle N(P) \rangle = N(P)$ by (1). Hence, $P \triangleleft G$ as required. \blacksquare

Note the proof of (2) shows that Sylow p -subgroups of $\Phi(G)$ are normal in G .

Corollary 1. Assume that $\Phi(G)$ is finitely generated (for example, if G is finite). If $H\Phi(G) = G$, where H is a subgroup, then $H = G$.

Proof. Write $\Phi(G) = \Phi$. If $\Phi = \langle t_1, \dots, t_k \rangle$, then $G = \langle H \cup \{t_1, \dots, t_k\} \rangle$ and the nongenerators t_i can be removed one by one. \blacksquare

The next result extends a useful theorem about finite p -groups.

Theorem 7. *If G is a nilpotent group and $\{1\} \neq H \triangleleft G$, then $H \cap Z(G) \neq \{1\}$.*

Proof. Write $\Gamma_i = \Gamma_i(G)$ for each i . We have $G = \Gamma_0 \supseteq \Gamma_1 \supseteq \dots \supseteq \Gamma_n = \{1\}$ for some n , so $H \cap \Gamma_n = \{1\}$ while $H \cap \Gamma_0 \neq \{1\}$. So there exists k such that $H \cap \Gamma_k \neq \{1\}$ while $H \cap \Gamma_{k+1} = \{1\}$. Choose $1 \neq h \in H \cap \Gamma_k$. If $g \in G$, then $[h, g] \in [\Gamma_k, G] = \Gamma_{k+1}$. Also, $[h, g] = h^{-1}g^{-1}hg \in H(g^{-1}Hg) = H$ because $H \triangleleft G$. So $[h, g] \in H \cap \Gamma_{k+1} = \{1\}$, whence $hg = gh$. Thus, $h \in Z(G) \cap H$. ■

In 1938, Hans Fitting identified a largest nilpotent normal subgroup in every finite group. His key result was

Theorem 8. Fitting's Theorem. *If H and K are nilpotent, normal subgroups of a finite group G , so also is HK .*

Proof. We proceed by induction on $|G|$. We have $HK \triangleleft G$, so we may assume (by induction) that $G = HK$ and that $H \neq \{1\} \neq K$. Write $W = Z(K)$. Then $W \neq \{1\}$ by Theorem 7. Also, $W \triangleleft G$ being characteristic in $K \triangleleft G$. If we write $N = [W, H]$, the proof falls into two cases.

Case 1. $N = \{1\}$. Then W centralizes H (and K), so $W \subseteq Z(HK) = Z(G)$. But $\frac{G}{W} = \frac{HW}{W} \frac{K}{W}$. Moreover, $\frac{HW}{W} \cong \frac{H}{H \cap W}$ and $\frac{K}{W}$ are both nilpotent by Theorem 1, so $\frac{G}{W}$ is nilpotent by induction. Hence, G is nilpotent by Theorem 2.

Case 2. $N \neq \{1\}$. We have $N \subseteq W \cap H$ because $W \triangleleft G$ and $H \triangleleft G$. In particular, $V = N \cap Z(H) \neq \{1\}$ again by Theorem 7. As before, $\frac{G}{V} = \frac{H}{V} \frac{KV}{V}$ and both $\frac{H}{V}$ and $\frac{KV}{V}$ are nilpotent, so $\frac{G}{V}$ is nilpotent by induction. But V centralizes H , and it also centralizes K because $V \subseteq N \subseteq W = Z(K)$. Hence, $V \subseteq Z(HK) = Z(G)$, so G is nilpotent by Theorem 2 and we are done in this case too. ■

Now let G be any finite group. If $N_1 = \{1\}$, N_2, \dots, N_k denote all the nilpotent, normal subgroups of G , define the **Fitting subgroup** $F(G)$ of G by

$$F(G) = \langle N_1 \cup N_2 \cup \dots \cup N_k \rangle = N_1 N_2 \dots N_k.$$

Then $F(G) \triangleleft G$, and it is nilpotent by Theorem 8 and induction on k . This proves

Theorem 9. *If G is a finite group, then $F(G)$ is the largest nilpotent, normal subgroup of G in the sense that it contains every such subgroup.*

Corollary 1. *If $\alpha : G \rightarrow H$ is an onto homomorphism, then $\alpha[F(G)] \subseteq F(H)$.*

Proof. $\alpha[F(G)] \triangleleft H$ because α is onto, and it is nilpotent by Theorem 1. ■

Hence, a finite group G is nilpotent if and only if $F(G) = G$. Clearly, $Z(G) \subseteq F(G)$, and $\Phi(G) \subseteq F(G)$ because it is nilpotent (Theorem 6) and normal in G (Theorem 5). Moreover, $F(G)$ is a characteristic subgroup of G by the above corollary.

Lemma 3. *If G is finite and $N \triangleleft G$, then N is nilpotent if and only if $N' \subseteq \Phi(G)$.*

Proof. Write $\Phi(G) = \Phi$. If $N' \subseteq \Phi$, then N' is nilpotent by Theorem 6, and so N is nilpotent by the corollary to Theorem 1. Conversely, if N is nilpotent, then $N' \subseteq \Phi(N)$ by Theorem 5, and it remains to show that $\Phi(N) \subseteq \Phi$. Write $\Phi(N) = H$

and suppose $H \not\subseteq \Phi$. Then $H \not\subseteq M$ for some maximal subgroup M of G , so $HM = G$. Since $H \subseteq N$, the modular law (Lemma 1 §8.1) gives

$$N = G \cap N = HM \cap N = H(M \cap N).$$

Thus, $N = \langle H \cup (M \cap N) \rangle$ and $H = \Phi(N)$ consist of (a finite number of) non-generators of N . Hence, $N = M \cap N$, whence $N \subseteq M$, a contradiction. ■

We can now describe the relationship between the Frattini and Fitting subgroups in a finite group.

Theorem 10. *Let G be a finite group and write $\Phi = \Phi(G)$ and $F = F(G)$.*

- (1) $F' \subseteq \Phi \subseteq F$.
- (2) $F(G/\Phi) = F/\Phi$.

Proof. (1) We have $F' \triangleleft G$ because it is characteristic in $F \triangleleft G$, so $F' \subseteq \Phi$ by Lemma 3 because F is nilpotent. On the other hand, $\Phi \subseteq F$ by Theorem 9 because Φ is nilpotent. This proves (1).

- (2) This follows from a more general result (Theorem 11).¹⁰⁸ ■

Theorem 11. *If G is a finite group and $\Phi(G) \subseteq N \triangleleft G$, then N is nilpotent if and only if $N/\Phi(G)$ is nilpotent.*

Proof. Write $\Phi(G) = \Phi$. If N is nilpotent, so is its image $\frac{N}{\Phi}$. For the converse, assume that $\frac{N}{\Phi}$ is nilpotent. To show that N is nilpotent, we show that every Sylow p -subgroup P of N is normal in N (and invoke Theorem 4). First, $\frac{\Phi P}{\Phi}$ is a Sylow p -subgroup of $\frac{N}{\Phi}$ (by Example 4 §8.4), whence $\frac{\Phi P}{\Phi} \triangleleft \frac{N}{\Phi}$ by hypothesis. But then $\frac{\Phi P}{\Phi}$ is characteristic in $\frac{N}{\Phi} \triangleleft \frac{G}{\Phi}$, and it follows that $\frac{\Phi P}{\Phi} \triangleleft \frac{G}{\Phi}$. Thus, $\Phi P \triangleleft G$. But P is a Sylow p -subgroup of ΦP (because $\Phi P \subseteq N$), so $G = (\Phi P)N_G(P)$ by Lemma 1 §8.4. Since G is finite and Φ consists of nongenerators, it follows that $G = P N_G(P) = N_G(P)$. Hence, $P \triangleleft G$, so certainly $P \triangleleft N$ as required. ■

There is much more information available on nilpotent groups in books on group theory.¹⁰⁹

Exercises 9.3

1. (a) Show that A_n is not nilpotent if $n \geq 3$.
 (b) Show that every nilpotent group is solvable, but not conversely.
2. Prove (1)–(4) in Lemma 1.
3. If H and K are subgroups of G and $\alpha: G \rightarrow G_1$ is a homomorphism, show that $\alpha[H, K] = [\alpha(H), \alpha(K)]$. Conclude that $[H, K]$ is normal in G (characteristic in G) if the same is true of H and K .

¹⁰⁸Because of (2) and the corollary to Theorem 9, the Fitting subgroup of G is also called the nilpotent radical of G .

¹⁰⁹The following books contain excellent introductions to the theory of nilpotent groups: MacDonald, I.D., *The Theory of Groups*, London: Oxford University Press, 1968; Rose, J.S., *A Course on Group Theory*, Cambridge, England: Cambridge University Press, 1978; Kargapolov, M.I. and Merzljakov, J.I., *Fundamentals of the Theory of Groups*, New York: Springer-Verlag, 1979; Gorenstein, D., *Finite Groups*, 2nd ed., New York: Chelsea, 1980.

4. If $\alpha : G \rightarrow H$ is any homomorphism, show that $\alpha[\Gamma_i(G)] = \Gamma_i[\alpha(G)]$.
5. If $G^{(k)}$ is the k th derived subgroup of G (see Theorem 1, Section 9.2), show that $G^{(k+1)} = [G^{(k)}, G^{(k)}]$ for each $k \geq 0$.
6. (a) Show that $\Gamma_i(G_1 \times \cdots \times G_n) \subseteq \Gamma_i(G_1) \times \cdots \times \Gamma_i(G_n)$ for all $i \geq 0$.
 (b) Show that equality holds in (a).
7. If $G = G_0 \supseteq G_1 \supseteq \cdots \supseteq G_n = \{1\}$ is any central series for a group G , show that $G_{n-i} \subseteq Z_i(G)$ for each i .
8. Let

$$G = \left\{ \begin{bmatrix} a & b \\ 0 & c \end{bmatrix} \mid a, b, c \in F; ac \neq 0 \right\},$$

- where F is a field. Is G nilpotent?
9. Show that D_n is nilpotent if and only if n is a power of 2.
 10. Show that a finite group is nilpotent if and only if any two elements of relatively prime orders commute.
 11. Show that a group G is nilpotent of class 2 if and only if G is nonabelian and $G' \subseteq Z(G)$.
 12. Show that a finite group G is nilpotent if and only if $Z(G/K)$ is nontrivial for all $K \triangleleft G$, $K \neq G$. [Hint: Theorem 7.]
 13. If G is a finite nilpotent group, let K be of minimal order in $\{K \mid \{1\} \neq K \triangleleft G\}$. Show that $K \subseteq Z(G)$ and that $|K|$ is a prime. [Hint: Theorem 7.]
 14. If $H \triangleleft G$ and $K \triangleleft G$, show that $G/(H \cap K)$ is nilpotent if and only if the product group $(G/H) \times (G/K)$ is nilpotent.
 15. Show that a finite group G is nilpotent if and only if G has a normal subgroup of order m for every divisor m of $|G|$.
 16. A subgroup H of a group G is called **subnormal** in G if a chain of subgroups $H = H_0 \subseteq H_1 \subseteq \cdots \subseteq H_n = G$ exists such that $H_i \triangleleft H_{i+1}$ for each i . Show that a finite group G is nilpotent if and only if every subgroup is subnormal.
 17. If $K \subseteq Z(G)$ and G/K is nilpotent, show that G is nilpotent using only (3) of Theorem 4.
 18. (a) If G is nilpotent, show that $Z(H) \neq \{1\}$ for all subgroups $H \neq \{1\}$.
 (b) Show that the converse is false by considering Q_6 .
 19. If G is nilpotent and G/G' is cyclic, show that G is abelian. [Hint: Apply Theorem 2 §2.9 to $G/\Gamma_2(G)$ and conclude that $\Gamma_2(G) = \Gamma_1(G)$.]
 20. If G is a finite group show that (2) \Leftrightarrow (3) \Leftrightarrow (4) and (2) \Rightarrow (1), but (1) $\not\Rightarrow$ (2).
 (1) G' is abelian. (2) $G/Z(G)$ is abelian. (3) $\Gamma_2(G) = \{1\}$. (4) $Z_2(G) = G$.
 21. Let $D_n = \langle a, b \rangle$ where $o(a) = n$, $o(b) = 2$, and $aba = b$. Show that (a) $\Phi(D_4) = \{1, a^2\}$, (b) $\Phi(D_{12}) \subseteq \{1, a^6\}$, and (c) $\Phi(D_{pq}) = \{1\}$, where $p \neq q$ are primes. [Hint: If $H \subseteq \langle a \rangle$ has index m , show that $\tilde{H} = H \cup Hb$ is a subgroup of index m .]
 22. If $o(a) = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, where p_i are distinct primes, show that $\Phi(\langle a \rangle) = \langle a^m \rangle$, where $m = p_1 p_2 \cdots p_r$.
 23. Let $|G| = p^3$, where p is a prime. If G is nonabelian, show that $\Phi(G) = G' = Z(G)$ and that this subgroup has order p . [Hint: Exercise 26 §8.2.]
 24. If G is a finite group, write $\Phi = \Phi(G)$. Show that the following are equivalent:
 (1) G is nilpotent (2) G/Φ is abelian (3) G/Φ is nilpotent
 25. If $K \triangleleft G$, where G is finite, and $\Phi(G/K) = \{K\}$, show that $\Phi(G) \subseteq K$.
 26. (a) If G is finite, $K \triangleleft G$, and $K \subseteq \Phi(G)$, show that $\Phi(G/K) = \Phi(G)/K$.
 (b) Show that $\Phi(G/\Phi(G)) = \{1\}$.

27. Show that $\Phi(G \times H) = \Phi(G) \times \Phi(H)$ for finite groups G and H .
28. If G is a finite group and $H \triangleleft G$, show that $\Phi(H) \subseteq H \cap \Phi(G)$. [Hint: If $\Phi(H) \not\subseteq M$, where M is maximal in G , show that $G = \Phi(H)M$ and apply Lemma 1 §8.1.]
29. (a) If G is a finite group and $M \subseteq G$ is a maximal subgroup, show that either $Z(G) \subseteq M$ or $G' \subseteq M$. [Hint: $MZ(G) = G$ implies that $M \triangleleft G$.]
(b) Show that $Z(G) \cap G' \subseteq \Phi(G)$ for all finite groups G .
30. Show that a finite group G can be generated by n elements if and only if the same is true of $G/\Phi(G)$.
31. If G is a finite p -group, p a prime, show that $\Phi(G) = \langle G' \cup \{g^p \mid g \in G\} \rangle$.

Chapter 10

Galois Theory

In most sciences, one generation tears down what another has built and what one has established another undoes. In mathematics alone, each generation adds a new storey to the old structure.

—Hermann Hankel

The moving power of mathematical invention is not reasoning but imagination.

—Augustus de Morgan

If $E \supseteq F$ is an extension of fields, Galois theory studies the set of automorphisms $\sigma : E \rightarrow E$ that fix F in the sense that $\sigma(a) = a$ for all $a \in F$. The set G of all such automorphisms is a group called the Galois group of E over F . With appropriate restrictions on the extension $E \supseteq F$, we can establish a bijection (called the Galois correspondence) between the subgroups of G and the subfields of E that contain F . This correspondence is very useful in deducing properties of the subfields from properties of the corresponding subgroups and conversely.¹¹⁰

The origins of Galois theory lie in the theory of equations. Methods implying the quadratic formula for solving $x^2 + bx + c = 0$ were known to the Babylonians in 1600 BC, but an algebraic formulation did not appear until the second century AD. As to cubics, nothing appears to have been done until the fifteenth century when Scipione del Ferro, and later Niccolò Tartaglia, found what is now called the cubic formula. This result, together with Lodovico Ferrari's formula for solving quartics, was published in 1545 in the book *Ars Magna* by the physician Girolamo Cardano.

¹¹⁰This chapter requires only Sections 6.1–6.4 as background. The material on solvable groups needed in Section 10.3 is adequately reviewed there.

After that the greatest mathematicians attempted to find a similar formula for expressing the roots of an arbitrary quintic in terms of coefficients by using only arithmetic operations and the extraction of n th roots (called radicals). Possibly the most important step was taken by Lagrange in 1770 when he unified the previous work by showing that, in every case, the solution depended on finding combinations of the roots of the equations that were unchanged when the roots were permuted. He showed that his method failed for the quintic, which aroused suspicion that a general formula was impossible in this case. A flawed proof of this impossibility by Ruffini appeared in 1813, and in 1824 Abel settled the matter once and for all: No general formula for the roots of a quintic exists that uses only radicals.

The general problem of determining which polynomial equations could be solved by radicals was resolved in 1830 by a 19-year-old Frenchman Évariste Galois. He had submitted three papers to the Academy of Sciences in Paris, but all were rejected. Incredibly, he was killed in a duel in 1832, and it was not until 1846 that his work finally received the recognition it deserved.

10.1 GALOIS GROUPS AND SEPARABILITY

If $E \supseteq F$ are fields, Galois theory is concerned with the automorphisms $\sigma: E \rightarrow E$ that fix F in the sense that $\sigma(a) = a$ for all $a \in F$. In this case, σ is called an **F -automorphism** of E . The identity automorphism ε certainly has this property, and one easily verifies that the set of F -automorphisms is a subgroup of the group of all automorphisms of E . This group is called the **Galois group** of the extension $E \supseteq F$ and is denoted $\text{gal}(E : F)$. We focus on this group throughout the chapter.

Example 1. $\text{gal}(F : F) = \{\varepsilon\}$ for all fields F .

Example 2. Show that $\text{gal}(\mathbb{C} : \mathbb{R}) = \{\varepsilon, \gamma\}$, where $\gamma: \mathbb{C} \rightarrow \mathbb{C}$ is the conjugation automorphism defined by $\gamma(z) = \bar{z}$ for all $z \in \mathbb{C}$.

Solution. If $\sigma \in \text{gal}(\mathbb{C} : \mathbb{R})$ and $z = a + bi$ in \mathbb{C} , then

$$\sigma(z) = \sigma(a + bi) = \sigma(a) + \sigma(b)\sigma(i) = a + b\sigma(i).$$

But $\sigma(i)^2 = \sigma(i^2) = \sigma(-1) = -1$, so $\sigma(i) = i$ or $\sigma(i) = -i$. These conditions give $\sigma = \varepsilon$ or $\sigma = \gamma$, respectively. \square

Throughout this chapter, we use terminology and notation for field extensions from Sections 6.1–6.4, usually without comment. For convenience, we restate three lemmas from Chapter 6 that will be referred to repeatedly. They come, respectively, from Theorem 4 §6.2, Theorem 3 §6.3, and Theorem 3 §6.2.

Lemma 1. If $E \supseteq F$ are fields, and $u \in E$ is algebraic over F of degree n , then:

- (1) $F(u) = \{f(u) \mid f \in F[x]\}$.
- (2) $\{1, u, \dots, u^{n-1}\}$ is an F -basis of $F(u)$.

If $\sigma: F \rightarrow \bar{F}$ is a ring homomorphism, and if $f = \sum a_i x^i \in F[x]$, recall that we define $f^\sigma \in \bar{F}[x]$ by $f^\sigma = \sum \sigma(a_i) x^i$.

Lemma 2. Let $\sigma : F \rightarrow \bar{F}$ be an isomorphism of fields F and \bar{F} and let p be a monic, irreducible polynomial in $F[x]$. If u and v are roots of p and p^σ in extension fields of F and \bar{F} , respectively, then there is an isomorphism

$$\hat{\sigma} : F(u) \rightarrow \bar{F}(v),$$

$$\begin{array}{ccc} F(u) & \xrightarrow{\hat{\sigma}} & \bar{F}(v) \\ \downarrow & & \downarrow \\ F & \xrightarrow{\sigma} & \bar{F} \end{array}$$

given by $\hat{\sigma}[f(u)] = f^\sigma(v)$ for each $f \in F[x]$, which extends σ .¹¹¹

Lemma 3. If $E \supseteq F$ is a field extension and $u \in E$ is algebraic over F , let m denote the monic polynomial of least degree such that $m(u) = 0$. Then m is uniquely determined by u , irreducible, and satisfies $f(u) = 0$, $f \in F[x]$, if and only if $m \mid f$.

The polynomial m in Lemma 3 is called the **minimal polynomial** of u over F , and the degree of m is called the **degree** of u over F and is written $\deg_F(u)$.

Let us return to Example 2. The fact that $\mathbb{C} = \mathbb{R}(i)$ is essential in the solution. Indeed \mathbb{C} contains all roots of the polynomial $x^2 + 1$, and the key observation in the solution is this: Given $\sigma \in \text{gal}(\mathbb{C} : \mathbb{R})$, the fact that i is a root of $x^2 + 1$ implies that $\sigma(i)$ is also a root. This basic fact is recorded as part of Lemma 4 (the proof is Exercise 2).

Lemma 4. If $E \supseteq F$ are fields, $G = \text{gal}(E : F)$, $u \in E$, and $\sigma \in G$, then

- (1) $\sigma[f(u)] = f[\sigma(u)]$ for all $f \in F[x]$.
- (2) In particular, if u is a root of f , then $\sigma(u)$ is also a root of f .
- (3) If u is algebraic over F , and $\sigma, \tau \in \text{gal}(F(u) : F)$, then $\sigma = \tau$ if and only if $\sigma(u) = \tau(u)$.

Let F be a field and let $F(u)$ be a simple extension of F where u is algebraic over F . We want to determine $\text{gal}(F(u) : F)$. By (3) of Lemma 4, each $\sigma \in \text{gal}(F(u) : F)$ is completely determined by the choice of $\sigma(u)$ in $F(u)$. But this choice is not arbitrary. If m is the minimal polynomial of u over F , then $m(u) = 0$, so $\sigma(u)$ is also a root of m by Lemma 4(2). Moreover, if $u_1 = u, u_2, \dots, u_r$ are the distinct roots of m in $F(u)$ then, by Lemma 2, an F -automorphism $\sigma_i : F(u) \rightarrow F(u)$ exists for each i such that $\sigma_i(u) = u_i$. Theorem 1 sums up this discussion.

Theorem 1. Let $F(u) \supseteq F$ be a simple extension, where u is algebraic over F with minimal polynomial $m \in F[x]$. If $u_1 = u, u_2, \dots, u_r$ are the distinct roots of m in $F(u)$, then

$$\text{gal}(F(u) : F) = \{\sigma_1 = \varepsilon, \sigma_2, \dots, \sigma_r\},$$

where, for each i , σ_i is the unique F -automorphism of $F(u)$ that satisfies $\sigma_i(u) = u_i$. Hence if m splits in $F(u)$ and has distinct roots then $|\text{gal}(F(u) : F)| = [F(u) : F]$.

Proof. It is clear that $\{\sigma_i \mid 1 \leq i \leq r\} \subseteq \text{gal}(F(u) : F)$, and the other inclusion follows from the above discussion. ■

Example 3. If $u = \sqrt[3]{2}$, show that $\text{gal}(\mathbb{Q}(u) : \mathbb{Q}) = \{\varepsilon\}$.

¹¹¹That is, $\hat{\sigma}(a) = \sigma(a)$ for all $a \in F$.

Solution. Here, $m = x^3 - 2$ is the minimal polynomial of u over \mathbb{Q} . The roots of m in \mathbb{C} are u, uw , and uw^2 , where $w = e^{2\pi i/3}$, so u is the only root in $\mathbb{Q}(u)$. Thus, any σ in $\text{gal}(\mathbb{Q}(u) : \mathbb{Q})$ must satisfy $\sigma(u) = u$, from which $\sigma = \varepsilon$. \square

Example 4. If $u = e^{2\pi i/5}$, write $G = \text{gal}(\mathbb{Q}(u) : \mathbb{Q})$. Show that $G \cong C_4$.

Solution. The roots in \mathbb{C} of $x^5 - 1$ are $1, u, u^2, u^3$, and u^4 , and they are distinct. Now $x^5 - 1 = (x - 1)\Phi_5(x)$, where $\Phi_5(x) = 1 + x + x^2 + x^3 + x^4$ is the fifth cyclotomic polynomial. This is \mathbb{Q} -irreducible by Example 13 §4.2 and so is the minimal polynomial of u . Hence, the roots of $\Phi_5(x)$ in \mathbb{C} are u, u^2, u^3 , and u^4 , and they are distinct and all lie in $\mathbb{Q}(u)$. It follows that $|G| = 4$ by Theorem 1.

By Lemma 2, $\sigma \in G$ exists with $\sigma(u) = u^2$. Then $\sigma^2(u) = \sigma(u^2) = \sigma(u)^2 = u^4$, so $\sigma^3(u) = \sigma(u^4) = u^8 = u^3$. Thus $\varepsilon, \sigma, \sigma^2$, and σ^3 are distinct by Lemma 4(3), so $G = \{\varepsilon, \sigma, \sigma^2, \sigma^3\}$. Clearly, $G \cong C_4$. \square

There is nothing special about the prime 5 in Example 4. Indeed, if p is any prime and $u = e^{2\pi i/p}$, the same argument shows that the minimal polynomial of u is $\Phi_p(x) = 1 + x + \cdots + x^{p-1}$ and that this polynomial has distinct roots u, u^2, \dots, u^{p-1} in $\mathbb{Q}(u)$. Hence, Theorem 1 shows that $G = \text{gal}[\mathbb{Q}(u) : \mathbb{Q}]$ satisfies $|G| = p - 1$. By Theorem 7 §6.4, \mathbb{Z}_p^* is cyclic, say $\mathbb{Z}_p^* = \langle \bar{m} \rangle = \{1, \bar{m}, \bar{m}^2, \dots, \bar{m}^{p-2}\}$. By Lemma 2, there exists $\sigma \in G$ satisfying $\sigma(u) = u^m$. Now let $\tau \in G$ be arbitrary. Then $\tau(u)$ is a root of $\Phi_p(x)$, say $\tau(u) = u^k$, where $1 \leq k \leq p - 1$. Thus, $k \equiv m^t \pmod{p}$, where $0 \leq t \leq p - 2$, so $u^k = u^{m^t}$ because $|u| = p$. In other words, $\tau(u) = \sigma^t(u)$, so $\tau = \sigma^t \in \langle \sigma \rangle$ by Lemma 4(3). Hence, $G \subseteq \langle \sigma \rangle$, so, since $|G| = p - 1$, this shows that $G = \langle \sigma \rangle \cong C_{p-1}$. We record this as Example 5.

Example 5. If p is a prime and $u = e^{2\pi i/p}$, then $\text{gal}(\mathbb{Q}(u) : \mathbb{Q}) \cong C_{p-1}$.

Example 6. Let $E = GF(p^n)$, where p is a prime, and regard \mathbb{Z}_p as a subfield of E . Show that $\text{gal}(E : \mathbb{Z}_p) \cong C_n$.

Solution. Write $G = \text{gal}(E : \mathbb{Z}_p)$. Corollary 1 of Theorem 7 §6.4 gives $E = \mathbb{Z}_p(u)$ for some $u \in E$, so $|G| \leq n$ by Theorem 1 because the minimal polynomial of u over \mathbb{Z}_p has degree $[E : \mathbb{Z}_p] = n$. On the other hand, let $\sigma : E \rightarrow E$ be the Frobenius automorphism defined by $\sigma(w) = w^p$ for all $w \in E$. Then $\sigma \in G$ by Fermat's theorem, and it suffices to show that $\sigma(u) = n$. One verifies that $\sigma^k(u) = u^{p^k}$ for all $k \geq 1$. Hence, if $\sigma^k = \varepsilon$, then every element of E is a root of $x^{p^k} - x$. As $|E| = p^n$, this condition implies that $k \geq n$. Hence $k = n$, as required. \square

Theorem 1 gives a lot of information about the Galois group of a simple algebraic extension, a situation that occurs commonly (see Theorem 6). However, many of the techniques used to prove it apply to any *finite* field extension $E \supseteq F$; that is, $[E : F] = \dim_F E$ is finite. Recall that $E \supseteq F$ is finite if and only if $E = F(u_1, u_2, \dots, u_n)$, where each $u_i \in E$ is algebraic over F (Theorem 6 §6.2).

Theorem 2. Let $E = F(u_1, \dots, u_n) \supseteq F$ be a finite extension where, for each i , u_i is algebraic over F with minimal polynomial m_i . If $\sigma \in \text{gal}(E : F)$, then

- (1) σ is uniquely determined by the choice of $\sigma(u_1), \dots, \sigma(u_n)$ in E .
- (2) $\sigma(u_i)$ is a root of m_i for each i .
- (3) If $\sigma, \tau \in \text{gal}(E : F)$, then $\sigma = \tau$ if and only if $\sigma(u_i) = \tau(u_i)$ for each i .

In particular, $\text{gal}(E : F)$ is a finite group.

Proof. (1) This follows from (3).

(2) We have $m_i[\sigma(u_i)] = \sigma[m_i(u_i)] = \sigma(0) = 0$ using Lemma 4.

(3) Let $\sigma(u_i) = \tau(u_i)$ for each i , where $\tau \in \text{gal}(E : F)$; we must show that $\sigma = \tau$. Writing $\lambda = \tau^{-1}\sigma$, it suffices to show the following: If $\lambda \in \text{gal}(E : F)$ satisfies $\lambda(u_i) = u_i$ for each i , then $\lambda = \varepsilon$. We prove this by induction on n . If $n = 1$, it is Lemma 4(3). If $n \geq 2$, write $K = F(u_1)$. Then λ fixes K , and it follows that $\lambda \in \text{gal}(K(u_2, \dots, u_n) : K)$. Thus, $\lambda = \varepsilon$ by induction. The last sentence follows from (1) and (2). \blacksquare

All the Galois groups that we have constructed so far are abelian. However, this is not the case in general; indeed, *every* finite group can be realized as a Galois group (Corollary 2 of Theorem 3 §10.3). For now, however, we content ourselves with constructing a nonabelian example using Theorem 2.

Example 7. If E is the splitting field¹¹² of $x^3 - 2$ over \mathbb{Q} , show that $\text{gal}(E : \mathbb{Q}) \cong D_3$.

Solution. Write $G = \text{gal}(E : \mathbb{Q})$, $u = \sqrt[3]{2}$, and $w = e^{2\pi i/3}$. Then the roots of $x^3 - 2$ are u, uw , and uw^2 , so $E = \mathbb{Q}(u, uw, uw^2) = \mathbb{Q}(u, w)$. The minimal polynomials of u and w over \mathbb{Q} are $x^3 - 2$ and $x^2 + x + 1$, respectively, and $x^2 + x + 1$ has roots w and w^2 in E . Thus, for $\sigma \in G$, Theorem 2 shows that $\sigma(u) \in \{u, uw, uw^2\}$ and $\sigma(w) \in \{w, w^2\}$, and hence that $|G| \leq 3 \times 2 = 6$. On the other hand, a \mathbb{Q} -isomorphism

$\sigma_0 : \mathbb{Q}(u) \rightarrow \mathbb{Q}(uw)$ exists by Lemma 2

with $\sigma_0(u) = uw$ (see Theorem 3 §6.3).

This isomorphism in turn extends (by the same theorem) to an automorphism σ of $E = \mathbb{Q}(u)(w) = \mathbb{Q}(uw)(w)$ with $\sigma(w) = w$ (see the figure). Thus, $\sigma \in G$ satisfies $\sigma(u) = uw$ and $\sigma(w) = w$. Similarly, $\tau \in G$ can be constructed such that $\tau(u) = u$ and $\tau(w) = w^2$. It is a routine matter (using

Theorem 2) to verify that $\sigma(\sigma) = 3$, $\sigma(\tau) = 2$, and $\sigma\tau\sigma = \tau$. Thus, $\langle \sigma, \tau \rangle \cong D_3$, so, because $|G| \leq 6$, $G = \langle \sigma, \tau \rangle$. \square

$$\begin{array}{ccc} E = \mathbb{Q}(u, w) & \xrightarrow{\sigma} & \mathbb{Q}(uw, w) = E \\ \downarrow & & \downarrow \\ \mathbb{Q}(u) & \xrightarrow{\sigma_0} & \mathbb{Q}(uw) \\ \downarrow & & \downarrow \\ \mathbb{Q} & \xrightarrow{\varepsilon} & \mathbb{Q} \end{array}$$

Separable Extensions

Let $G = \text{gal}(E : F)$, where E is the splitting field of a polynomial f over F , and let X denote the set of distinct roots of f in E . If $\sigma \in G$, then $\sigma(u) \in X$ for all $u \in X$, so we have the *restriction* map $\sigma|_X : X \rightarrow X$ defined by

$$\sigma|_X(u) = \sigma(u) \quad \text{for all } u \in X.$$

Then $\sigma|_X \in S_X$ because σ is one-to-one and X is finite, and $\sigma \mapsto \sigma|_X$ is a group homomorphism that is one-to-one by Theorem 2. Hence, we can view G as a group of permutations of X . The following terminology is standard. A group G of permu-

¹¹²Splitting fields are discussed in detail in Section 6.3.

tations of a set X is said to **act transitively** on X if, for all $u, v \in X$, there exists $\sigma \in G$ such that $\sigma(u) = v$.

Theorem 3. Let $G = \text{gal}(E : F)$, where E is the splitting field of a polynomial f over F , and let X denote the set of distinct roots of f in E . Then

- (1) G is isomorphic (by restriction) to a subgroup of S_X .
- (2) If f is irreducible in $F[x]$, then G acts transitively on X .
- (3) If f has no repeated root in E , and G acts transitively on X , then f is irreducible in $F[x]$.

Proof. (1) This follows by the discussion preceding this theorem.

(2) If $u, v \in X$, then E is the splitting field of f over $F(u)$ and also over $F(v)$. Hence, Lemma 2 gives an F -isomorphism $\sigma_0 : F(u) \rightarrow F(v)$ with $\sigma_0(u) = v$. This isomorphism extends to $\sigma \in G$ by Theorem 4 §6.3.

(3) Suppose that $f = gh$ in $F[x]$; g, h not constant. Let $g(u) = 0 = h(v)$, where $u, v \in X$. Because G acts transitively, let $v = \sigma(u)$, where $\sigma \in G$. Then $g(v) = g[\sigma(u)] = \sigma[g(u)] = 0$, so v is a repeated root of f , contrary to hypothesis. ■

If $E \supseteq F$ is a finite extension of fields, we want to determine the size of the Galois group $G = \text{gal}(E : F)$. If $E = F(u)$, where $u \in E$ is algebraic over F , Theorem 1 shows that $|G|$ is the number of distinct roots in E of the minimal polynomial of u . If $E = F(u_1, \dots, u_n)$ and m_i is the minimal polynomial of u_i for each i , Theorem 2 shows that $\sigma \in G$ is determined by its effect on the roots of these polynomials m_i . To count these automorphisms, we adopt a different perspective.

We assume that E is the splitting field¹¹³ of a polynomial f in $F[x]$. We are going to prove that if every irreducible factor of f has distinct roots in E , the Galois group $G = \text{gal}(E : F)$ has order $|G| = [E : F]$. Examples 5 and 7 illustrate this. The next result provides a simple test for when an irreducible polynomial has distinct roots. The test involves the formal derivative f' of a polynomial f , defined in Section 6.4 as follows: If

$$f = a_0 + a_1x + \dots + a_nx^n, \quad \text{then} \quad f' = a_1 + 2a_2x + \dots + na_nx^{n-1}.$$

The usual properties of derivatives remain valid (Theorem 2 §6.4).

Lemma 5. If F is a field, the following conditions are equivalent for an irreducible polynomial p in $F[x]$.

- (1) p has distinct roots in every extension field of F in which it splits.
- (2) p has distinct roots in some splitting field of p over F .
- (3) $p' \neq 0$.

Proof. (1) \Rightarrow (2). This is clear.

(2) \Rightarrow (3). Let $E \supseteq F$ be a splitting field for p over F and let $p(u) = 0$, $u \in E$. If $p' = 0$, then $x - u$ divides both p and p' and so $(x - u)^2$ divides p by Theorem 3 §6.4, contrary to (2). So $p' \neq 0$.

(3) \Rightarrow (1). Suppose that p splits in $E \supseteq F$ and assume that $u \in E$ is a repeated root of p in E . Then $(x - u)^2$ divides p in $E[x]$ and so $(x - u)$ divides both p and

¹¹³Not every finite extension is a splitting field. For example, $\mathbb{Q}(\sqrt[3]{2})$ is not a splitting field of any polynomial in $\mathbb{Q}[x]$. We discuss this topic further in Section 10.2.

p' in $E[x]$ by Theorem 3 §6.4. But p and p' are relatively prime— p is irreducible and does not divide p' , so $(x - u)$ divides 1 in $E[x]$, a contradiction. ■

If F is a field, an irreducible polynomial p in $F[x]$ is called **separable** over F if it satisfies the conditions in Lemma 5, and a polynomial $f \in F[x]$ of positive degree is called **separable** over F (or separable in $F[x]$) if all its irreducible factors are separable. An extension $E \supseteq F$ of fields is called a **separable extension** if it is algebraic and the minimal polynomial of each element of E is separable over F .

Example 8. The irreducible polynomial $p = x^2 + 2$ is separable over \mathbb{Q} because $p' \neq 0$, or because its roots $\pm i\sqrt{2}$ in $\mathbb{C} \supseteq \mathbb{Q}$ are distinct. Hence, the polynomial $x^4 + 4x^2 + 4 = (x^2 + 2)^2$ is also separable over \mathbb{Q} .

Example 9. Show that $f = x^6 - x^3 - 1$ is separable over \mathbb{Z}_3 . However, $f' = 0$.

Solution. We have $f = p^3$, where $p = x^2 - x - 1$ is irreducible over \mathbb{Z}_3 . Hence, it suffices to show that p is separable. But p is separable by Lemma 5 because $p' = 2x - 1 \neq 0$. However, $f' = 0$ because $\text{char } \mathbb{Z}_3 = 3$. □

Let $f = a_0 + a_1x + a_2x^2 + \cdots + a_kx^k + \cdots$ be a polynomial in $F[x]$. When the formal derivative $f' = 0$ depends on the characteristic of the field F , we have

$$f' = a_1 + 2a_2x + \cdots + ka_kx^{k-1} + \cdots$$

so $f' = 0$ if and only if $ka_k = 0$ for all $k \geq 1$. If $\text{char } F = 0$, this implies that $a_k = 0$ for all $k \geq 1$, so $f = a_0$ is constant (as in calculus). However, if $\text{char } F = p$ is a prime, then $f' = 0$ implies that $a_k = 0$ whenever p does not divide k , that is, when $f = g(x^p)$ for some polynomial g in $F[x]$. Conversely, Theorem 2 §6.4 gives $[g(x^p)]' = g'(x^p)(px^{p-1}) = 0$ when the characteristic is p . With Lemma 5, this observation gives Theorem 4.

Theorem 4. Let f be an irreducible polynomial in $F[x]$, where F is a field.

- (1) If $\text{char } F = 0$, then f is separable over F .
- (2) If $\text{char } F = p$, then f is separable over F if and only if it is not of the form $f = g(x^p)$ for some polynomial $g \in F[x]$.

Corollary. If $\text{char } F = 0$, every algebraic extension of F is separable.

Our goal is to show that if $E \supseteq F$ is the splitting field of a separable polynomial in $F[x]$, then the Galois group has order $[E : F]$. It is convenient to prove slightly more. Suppose that $\sigma : F \rightarrow \bar{F}$ is an isomorphism of fields. If $f = \sum_{i=0}^n a_i x^i$ is a polynomial in $F[x]$, recall that $f^\sigma \in \bar{F}[x]$ is defined by $f^\sigma(x) = \sum_{i=0}^n \sigma(a_i)x^i$. If $E \supseteq F$ and $\bar{E} \supseteq \bar{F}$ are splitting fields of f and f^σ , respectively, then Theorem 4 §6.3 asserts that an isomorphism $\hat{\sigma} : E \rightarrow \bar{E}$ exists that extends σ (that is, $\hat{\sigma}(a) = \sigma(a)$ for all $a \in F$). If f is a separable polynomial, we can count such extensions.

Theorem 5. Let $\sigma : F \rightarrow \bar{F}$ be an isomorphism of fields and let $f \in F[x]$ is a separable polynomial. If $E \supseteq F$ and $\bar{E} \supseteq \bar{F}$ are splitting fields of f and f^σ , respectively, there are exactly $[E : F]$ isomorphisms $\hat{\sigma} : E \rightarrow \bar{E}$ that extend σ .

Proof. Use induction on $[E : F]$. If $[E : F] = 1$, then $E = F$ and f splits in $F[x]$; that is, $f = a(x - a_1) \cdots (x - a_n)$, where $a, a_i \in F$. Since $f \mapsto f^\sigma$ is a ring

homomorphism, $f^\sigma = \sigma(a)(x - \sigma(a_1)) \cdots (x - \sigma(a_n))$ splits in \bar{F} . This means that $\bar{E} = \bar{F}$ and the only extension is $\hat{\sigma} = \sigma$.

If $[E : F] > 1$, then f does not split in $F[x]$, so let p be an irreducible factor of f with $\deg p = k \geq 2$. Fix a root $u \in E$ of p . Then any isomorphism $\hat{\sigma} : E \rightarrow \bar{E}$ induces an isomorphism $\tau : F(u) \rightarrow K$, where $K = \hat{\sigma}[F(u)]$ is a subfield of \bar{E} containing \bar{F} (see the diagram). Obviously, $\hat{\sigma}$ extends τ and τ extends σ . Hence, the number of possibilities for $\hat{\sigma}$ equals the number of extensions τ of σ times the number of extensions $\hat{\sigma}$ of τ . Now the multiplication theorem gives

$$[E : F(u)] = \frac{[E : F]}{[F(u) : F]} = \frac{[E : F]}{k} < [E : F].$$

Moreover, E is the splitting field of f over $F(u)$, and f remains separable over $F(u)$ because any irreducible factor of f in $F(u)[x]$ must divide an irreducible factor of f in $F[x]$. Hence, by induction, the number of extensions of τ to E is $[E : F(u)] = [E : F]/k$. So it remains to show that there are exactly k one-to-one ring homomorphisms $\tau : F(u) \rightarrow \bar{E}$ that extend σ .

But $f \mapsto f^\sigma$ is a ring isomorphism $F[x] \rightarrow \bar{F}[x]$, so p^σ is irreducible of degree k in $\bar{F}[x]$. Moreover, $p' \neq 0$ (f is separable by hypothesis) so $(p^\sigma)' = (p')^\sigma \neq 0$. Thus, p^σ has m distinct roots v_1, \dots, v_m in \bar{E} , and Theorem 3 §6.3 shows that, for each i , an isomorphism $\tau_i : F(u) \rightarrow \bar{F}(v_i)$ exists that extends σ and satisfies $\tau_i(u) = v_i$. Hence, $\{\tau_1, \tau_2, \dots, \tau_m\}$ are distinct extensions of σ to $F(u)$. But if τ is any such extension, then $p^\sigma[\tau(u)] = \tau[p(u)] = \tau(0) = 0$, so $\tau(u) = v_i = \tau_i(u)$ for some i . Hence, $\tau = \tau_i$, which completes the proof. ■

If we take $F = \bar{F}$, $E = \bar{E}$, and $\sigma = \varepsilon$ in Theorem 5, we obtain

Corollary. Let $E \supseteq F$ be a splitting field of a separable polynomial in $F[x]$. If $G = \text{gal}(E : F)$, then $|G| = [E : F]$.

The corollary will be used several times below. In fact, extensions $E \supseteq F$, where E is the splitting field of a separable polynomial over F , will occupy much of our effort in Section 10.2. We conclude this section with the surprising fact that every finite separable extension is simple.

Theorem 6. Primitive Element Theorem. Let $E \supseteq F$ be a finite separable extension. Then E is a simple extension of F ; that is, $E = F(u)$ for some $u \in E$.

Proof. If F is a finite field, then E is also finite, so the unit group E^* is cyclic by Theorem 7 §6.4, say $E^* = \langle u \rangle$. Hence, $E = F(u)$.

So assume that F is infinite. By Theorem 6 §6.2 (and induction), we may assume that $E = F(v, w)$. Let p and q be the minimal polynomials over F for v and w , and let $v_1 = v, v_2, \dots, v_m$ and $w_1 = w, w_2, \dots, w_n$, respectively, be the roots of p and q in E . The variables v_i are distinct because p is separable (take a splitting field of p containing E). Similarly, w_j are distinct. As F is infinite, $a \in F$ exists such that

$$a \neq \frac{v_i - v}{w - w_j} \quad \text{for all } i \text{ and all } j \neq 1.$$

$$\begin{array}{ccc} E & \xrightarrow{\hat{\sigma}} & \bar{E} \\ \downarrow & & \downarrow \\ F(u) & \xrightarrow{\tau} & K \\ \downarrow & & \downarrow \\ F & \xrightarrow{\sigma} & \bar{F} \end{array}$$

If $u = v + aw$, then $F(u) \subseteq F(v, w) = E$, and we claim this is equality. For this it suffices to show that $w \in F(u)$. Write $K = F(u)$ for convenience, and let m be the minimal polynomial of w over K . Then it suffices to show that $w \in K$ or, equivalently, that m is linear. Now $m|q$ (because $q(w) = 0$), so m is the product of some of the factors $x - w_j$. On the other hand, define $f = p(u - ax) \in K[x]$. Then $f(w) = p(v) = 0$, so $m|f$. However, $f(w_j) \neq 0$ for all $j \neq 1$ by the choice of a and u , so $m(w_j) \neq 0$. Thus $m = x - w$, as required. ■

Corollary. If F has characteristic 0, any finite extension of F is simple.

The proof of the primitive element theorem actually gives an algorithm for finding a generator of the extension. Here is an example.

Example 10. Let $F = \mathbb{Q}$ and $E = \mathbb{Q}(\sqrt{2}, \sqrt{5})$. In the notation of the proof of Theorem 6, we write $v = \sqrt{2}$ and $w = \sqrt{5}$, so the minimal polynomials are $p = x^2 - 2$ and $q = x^2 - 5$. Then $v_1 = \sqrt{2}$, $v_2 = -\sqrt{2}$, $w_1 = \sqrt{5}$, and $w_2 = -\sqrt{5}$, so the quantities $(v_i - v)/(w - w_j)$ in the proof reduce to 0 and $-\sqrt{2}/\sqrt{5}$. If we choose $a = 1$, the proof gives $E = \mathbb{Q}(\sqrt{2} + \sqrt{5})$, as we showed directly in Example 15 §6.2.

Exercises 10.1

Throughout these exercises, E and F are assumed to be fields.

1. Prove that $\text{gal}(E : F)$ is a group for any field extension $E \supseteq F$.
2. Prove Lemma 4.
3. If $E \supseteq F$ and $\{u_1, \dots, u_n\}$ is an F -basis of E , show that $\sigma \in \text{gal}(E : F)$ is uniquely determined by the choice of $\sigma(u_1), \dots, \sigma(u_n)$.
4. If $E \supseteq F$ and $u \in E$, show that $\text{gal}(E : F(u)) = \{\sigma \in \text{gal}(E : F) \mid \sigma(u) = u\}$.
5. If $E \supseteq \mathbb{Q}$, show that $\text{gal}(E : \mathbb{Q}) = \text{aut } E$.
6. If $E = \mathbb{Q}(e^{2\pi i/8})$, compute $\text{gal}(E : \mathbb{Q})$.
7. If $E = \mathbb{Q}(e^{2\pi i/6})$, compute $\text{gal}(E : \mathbb{Q})$.
8. If $E = \mathbb{Q}(\sqrt{2}, \sqrt{3})$, show that $\text{gal}(E : \mathbb{Q}) \cong C_2 \times C_2$. [Hint: Lemma 2.]
9. If $E = \mathbb{Q}(i, \sqrt{3})$, compute $\text{gal}(E : \mathbb{Q})$.
10. (a) If $E = \mathbb{Q}(\sqrt[4]{2})$, show that $\text{gal}(E : \mathbb{Q}) \cong C_2$. [Hint: Lemma 2.]
(b) Why does (a) not contradict the corollary to Theorem 5?
11. If $[E : F] = 2$, show that $\text{gal}(E : F) \cong C_2$.
12. Let E be the splitting field of $f = x^6 + x^3 - 1$ over \mathbb{Z}_3 . Show that $E = \mathbb{Z}_3(u)$ is a simple extension and find $\text{gal}(E : \mathbb{Z}_3)$. [Hint: $(a+b+c)^3 = a^3 + b^3 + c^3$ in \mathbb{Z}_3 .]
13. If $E = \mathbb{Q}(\sqrt[4]{2}, i)$, show that $\text{gal}(E : \mathbb{Q}) \cong D_4$. [Hint: If $u = \sqrt[4]{2}$, find σ and τ in $\text{gal}(E : \mathbb{Q})$ such that $\sigma(u) = iu$, $\sigma(i) = i$, $\tau(u) = u$, and $\tau(i) = -i$.]
14. Let E be the splitting field over \mathbb{Q} of $x^n - 1$. Show that $\text{gal}(E : \mathbb{Q})$ is abelian.
15. Use the method of Example 10 to show that $E = \mathbb{Q}(u)$ if
 - (a) $E = \mathbb{Q}(\sqrt{3}, \sqrt{5})$
 - (b) $E = \mathbb{Q}(i, \sqrt{5})$
16. (a) Show that $\mathbb{Q}(\sqrt{p}, \sqrt{q}) = \mathbb{Q}(\sqrt{p} + \sqrt{q})$, where p and q are distinct primes. [Hint: Example 10.]
(b) Show that $\mathbb{Q}(\sqrt{p}, \sqrt{q}, \sqrt{r}) = \mathbb{Q}(\sqrt{p} + \sqrt{q} + \sqrt{r})$, where p, q and r are distinct primes. [Hint: Exercise 32 §6.2.]

17. Show that $\text{gal}(\mathbb{R} : \mathbb{Q}) = \{\varepsilon\}$. [Hint: If $u < v$ in \mathbb{R} , show that $\sigma(u) < \sigma(v)$ for all σ in $\text{gal}(\mathbb{R} : \mathbb{Q})$ because $v - u = w^2$, $w \in \mathbb{R}$. If $u < \sigma(u)$, choose $a \in \mathbb{Q}$ such that $u < a < \sigma(u)$.]
18. Let $u = e^{\pi i/q}$, where q is an odd prime. Show that $\text{gal}(\mathbb{Q}(u) : \mathbb{Q}) \cong C_{q-1}$. [Hint: Show that $\mathbb{Q}(u^2) = \mathbb{Q}(u)$.]
19. Let $E = F(u_1, u_2, \dots, u_n)$, where each u_i is algebraic over F . If $\sigma, \tau \in \text{gal}(E : F)$ satisfy $\sigma(u_i) = \tau(u_i)$ for each i , show that $\sigma = \tau$.
20. Let $F = K(t)$ denote the field of rational forms over a field K in an indeterminate t . Show that $x^2 - t$ is irreducible over F but is not separable if $\text{char } K = 2$.
21. Let $F(t)$ denote the field of rational forms over a field F . Given $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in $GL_2(F)$, define $\sigma_M : F(t) \rightarrow F(t)$ by $\sigma_M[\lambda(t)] = \lambda \left(\frac{at+b}{ct+d} \right)$. Show that $M \mapsto \sigma_M$ is an onto group homomorphism $GL_2(F) \rightarrow \text{gal}[F(t) : F]$ with kernel

$$Z[GL_2(F)] = \left\{ \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \mid 0 \neq a \in F \right\}.$$

22. (a) Show that the following are equivalent for a polynomial f in $F[x]$.
- (1) f has no repeated root in any extension field of F .
 - (2) f has no repeated root in some splitting field over F .
 - (3) f and f' are relatively prime in $F[x]$.
- (b) If f is as in (a), show that f is separable, but not conversely.
23. If $n \geq 2$, show that $f = x^n - x \in F[x]$ has no repeated root in any splitting field if either $\text{char } F = 0$ or $\text{char } F = p$ and p does not divide $n - 1$. [Hint: Exercise 22.]
24. If $\text{char } F = p$ and F contains n distinct n th roots of unity, show that p does not divide n . [Hint: Exercise 22.]
25. If $E \supseteq F$ and $f \in F[x]$ is separable over F , show that f is separable over E .
26. If $E \supseteq K \supseteq F$ and $E \supseteq F$ is a separable extension, show that both $E \supseteq K$ and $K \supseteq F$ are separable extensions. [Remark: The converse is true if $[E : F]$ is finite—see Exercise 31.]
27. Let F have characteristic p . If $f = x^p - a$, where $a \in F$, show that f is irreducible or a power of a linear polynomial. [Hint: Lemma 5 and Theorem 4.]
28. (a) Show that the following are equivalent for F (then called a **perfect field**):
- (1) Every algebraic extension of F is separable.
 - (2) Every finite extension of F is separable.
 - (3) Every irreducible polynomial in $F[x]$ is separable.
- (b) Show that every field of characteristic 0 is perfect.
- (c) Show that every algebraic extension of a perfect field is perfect.
29. (a) Let F be a field of characteristic p . Show that F is perfect (Exercise 28) if and only if every element $b \in F$ has the form $b = a^p$ for some $a \in F$. [Hint: If F is perfect and $a \in F$, consider the irreducible factors of $x^p - a$ in some splitting field. For the converse, use Theorem 4.]
- (b) Show that every finite field is perfect.
30. Let $E \supseteq F$ be a finite extension, where $\text{char } F = p$.
- (a) If $u \in E$ has a separable minimal polynomial q over F , show that $u \in F(u^p)$. [Hint: If m is the minimal polynomial of u over $F(u^p)$, show $m|q$ and $m|(x - u)^p$.]
 - (b) Define $F(E^p) = \{a_1 u_1^p + \dots + a_n u_n^p \mid a_i \in F, u_i \in E, n \geq 1\}$. Show that $F(E^p)$ is a subfield of E . [Hint: Exercise 35 §6.2.]

- (c) If $E = F(E^p)$ and $\{w_1, \dots, w_k\} \subseteq E$ is F -independent, show that $\{w_1^p, \dots, w_k^p\}$ is F -independent. [Hint: Extend to a basis $\{w_1, \dots, w_k, \dots, w_n\}$ of E , show that $\{w_1^p, \dots, w_k^p, \dots, w_n^p\}$ spans E , and apply Theorem 7 §6.1.]
- (d) Show that $E \supseteq F$ is separable if and only if $F(E^p) = E$. [Hint: If $E = F(E^p)$, use Theorem 4 §6.2 and (c).]
31. Let $E \supseteq K \supseteq F$ be fields with $[E : F]$ finite. Show that $E \supseteq F$ is separable if and only if both $E \supseteq K$ and $K \supseteq F$ are separable. [Hint: Exercises 26 and 30.]
32. If $E \supseteq F$ is a finite extension, then $u \in E$ is called a **separable element** over F if its minimal polynomial in $F[x]$ is separable.
- If $u \in E$ is separable over F and $E \supseteq K \supseteq F$, where K is a field, show that u is separable over K . [Hint: Exercise 30(d).]
 - Show that $u \in E$ is separable over F if and only if $F(u) \supseteq F$ is a separable extension.
 - Define $S = \{u \in E \mid u \text{ is separable over } F\}$. Show that S is a subfield of E , that $S \supseteq F$ is separable, and that $E \supseteq K \supseteq F$, with $K \supseteq F$ separable, implies that $S \supseteq K$. The field S is called the **separable closure** of F in E . [Hint: If $u, v \in S$, show that $F(u, v) \supseteq F$ is separable by (a) and Exercise 31.]

10.2 THE MAIN THEOREM OF GALOIS THEORY

The central theme of Galois theory is to analyze a field extension $E \supseteq F$ by studying its Galois group $G = \text{gal}(E : F)$. It turns out that a beautiful correspondence exists between the subgroups H of G and the intermediate fields K with $E \supseteq K \supseteq F$. This correspondence was first noticed by Galois in his study of the roots of polynomials, published in 1846, but it was not until 1894 that Richard Dedekind first formulated the theory in terms of field extensions. We begin with two of Dedekind's theorems on field automorphisms in the form given in 1942 by Emil Artin in his definitive account of the subject.¹¹⁴ The first of these results is more general than needed here, but the additional generality involves little extra effort, improves the exposition, and introduces the concept of a group character, which is important in the theory of group representations.

Let G be a group and let E be a field. A group homomorphism $\sigma : G \rightarrow E^*$ is called a **character** of G in E . A set $\{\sigma_1, \dots, \sigma_n\}$ of characters of G in E is called **independent**¹¹⁵ if, given u_1, \dots, u_n in E ,

$$u_1 \sigma_1(g) + u_2 \sigma_2(g) + \cdots + u_n \sigma_n(g) = 0 \text{ for all } g \in G$$

implies that $u_1 = u_2 = \cdots = u_n = 0$.

Lemma 1. Dedekind's Lemma. Let $\{\sigma_1, \dots, \sigma_n\}$ be a finite set of distinct characters of a group G in a field E . Then $\{\sigma_1, \dots, \sigma_n\}$ is independent.

¹¹⁴Artin, E., *Galois Theory*, 2nd ed., Notre Dame Mathematical Lectures No. 2, University of Notre Dame, 1964.

¹¹⁵This property is independent in the vector space V of all mappings $\sigma : G \rightarrow E$, where addition and scalar multiplication are defined by $(\sigma + \tau)(g) = \sigma(g) + \tau(g)$ and $(u\sigma)(g) = u\sigma(g)$ for all $g \in G$, all $\sigma, \tau \in V$, and all $u \in E$.

Proof. For simplicity, write $\sigma_i(g) = \sigma_i g$ for each i and all $g \in G$. Proceed by induction on the number n of distinct characters. If $n = 1$, then $u_1 \sigma_1 g = 0$ for all $g \in G$ implies that $u_1 = 0$ because $\sigma_1 g \neq 0$. If $n > 1$, assume that

$$u_1 \sigma_1 g + u_2 \sigma_2 g + \cdots + u_n \sigma_n g = 0 \quad \text{for all } g \in G. \quad (*)$$

We must show that $u_i = 0$ for all i . If not, we may assume (by induction) that $u_i \neq 0$ for all i . Given $h \in G$, replace g by gh in $(*)$ and use the fact that σ_i are homomorphisms to get

$$u_1 \sigma_1 g \sigma_1 h + u_2 \sigma_2 g \sigma_2 h + \cdots + u_n \sigma_n g \sigma_n h = 0 \quad \text{for all } g \in G. \quad (**)$$

If $(*)$ is multiplied by $\sigma_1 h$ and the result is subtracted from $(**)$, the first terms cancel and the result is

$$u_2(\sigma_2 h - \sigma_1 h) \sigma_2 g + \cdots + u_n(\sigma_n h - \sigma_1 h) \sigma_n g = 0 \quad \text{for all } g \in G.$$

Thus, $u_i(\sigma_i h - \sigma_1 h) = 0$ for all $i \geq 2$ by induction, which yields $\sigma_i h = \sigma_1 h$ because $u_i \neq 0$. Because this is true for all $h \in G$, it implies that $\sigma_i = \sigma_1$ for each i , contrary to the hypothesis that they are distinct. \blacksquare

For Galois theory, the most interesting use of Lemma 1 arises as follows: If $\sigma : E \rightarrow E$ is an automorphism of the field E , the restriction of σ to the group E^* of units of E is a group homomorphism $E^* \rightarrow E^*$ and so is a character of E^* in E . This gives the following corollary .

Corollary. Any finite set of automorphisms of a field E is independent.

If $E \supseteq F$ is the splitting field of a separable polynomial in $F[x]$, then the corollary to Theorem 5 §10.1 gives $|\text{gal}(E : F)| = [E : F]$. Dedekind's lemma gives us half of this for any finite extension.

Theorem 1. Let $E \supseteq F$ be a finite extension of fields. If $G = \text{gal}(E : F)$ denotes the Galois group, then

$$|G| \leq [E : F].$$

Proof. Write $[E : F] = n$ and let $\{v_1, \dots, v_n\}$ be an F -basis of E ; we must show that $|G| \leq n$. If $|G| > n$, let $\sigma_0, \sigma_1, \dots, \sigma_n$ be distinct elements of G and write $\sigma_i(g) = \sigma_i g$ for $g \in G$ as before. Since each σ_i fixes F , it is F -linear in the sense that $\sigma_i(av) = a\sigma_i(v)$ for all $a \in F$ and $v \in E$.

Now consider the following set of n equations in $n+1$ variables x_0, x_1, \dots, x_n :

$$\sigma_0 v_1 x_0 + \sigma_1 v_1 x_1 + \cdots + \sigma_n v_1 x_n = 0,$$

$$\sigma_0 v_2 x_0 + \sigma_1 v_2 x_1 + \cdots + \sigma_n v_2 x_n = 0,$$

$$\vdots \quad \vdots \quad \vdots$$

$$\sigma_0 v_n x_0 + \sigma_1 v_n x_1 + \cdots + \sigma_n v_n x_n = 0.$$

Because there are more variables than equations, a solution $x_j = u_j \in E$ exists where $u_j \neq 0$ for some j . Thus,

$$\sum_{j=0}^n \sigma_j v_i u_j = 0 \quad \text{for } i = 1, 2, \dots, n.$$

Given $u \in E$, write $u = \sum_{i=1}^n a_i v_i$, $a_i \in F$. Since each σ_j is F -linear, we get

$$\sum_{j=0}^n u_j \sigma_j u = \sum_{j=0}^n u_j \left(\sum_{i=1}^n a_i \sigma_j v_i \right) = \sum_{i=1}^n a_i \left(\sum_{j=0}^n u_j \sigma_j v_i \right) = \sum_{i=1}^n a_i 0 = 0.$$

This is a contradiction because the σ_j are independent by the corollary to Dedekind's lemma. ■

No algebraist could resist the temptation to discover when equality holds in Theorem 1. To study this question, we need a concept that reflects Artin's point of view that the basic object of study in Galois theory is a field E , together with a group G of automorphisms of E . In this case, write

$$E_G = \{u \in E \mid \sigma(u) = u \text{ for all } \sigma \in G\} = \{u \in E \mid G \text{ fixes } u\}.$$

It is easy to verify that E_G is a subfield of E , called the **fixed field** of G in E . Note that $G \subseteq \text{gal}(E : E_G)$. If G is finite, we have the following fundamental result that, although stated originally by Dedekind, has become known as the Dedekind–Artin theorem.

Theorem 2. Dedekind–Artin Theorem. *Let E be field and let G be a finite group of automorphisms of E . Then $[E : E_G]$ is finite and*

$$[E : E_G] = |G|.$$

Proof. Write $E_G = F$ and $|G| = n$. If $[E : F]$ is finite, then $n \leq [E : F]$ by Theorem 1 because $G \subseteq \text{gal}(E : F)$. Hence, the proof is completed by showing that $n < [E : F]$ leads to a contradiction. In this case, let $\{u_0, u_1, \dots, u_n\} \subseteq E$ be independent over F . Consider the following set of $|G| = n$ equations in $n + 1$ variables x_0, x_1, \dots, x_n where, once again, we write $\sigma(u_j) = \sigma u_j$ whenever $\sigma \in G$.

$$\sigma u_0 x_0 + \sigma u_1 x_1 + \dots + \sigma u_n x_n = 0, \quad \sigma \in G. \quad (*)$$

Because there are more variables than equations, there is a solution with not all variables zero. Among all such solutions, choose one with the smallest number $r + 1$ of nonzero values. By relabeling variables if necessary, we may assume that $x_0 = v_0, \dots, x_r = v_r$ are these nonzero values and (multiplying by v_0^{-1}) assume further that $v_0 = 1$. Then $(*)$ becomes

$$\sigma u_0 + \sigma u_1 v_1 + \dots + \sigma u_r v_r = 0, \quad \sigma \in G. \quad (**)$$

Taking $\sigma = \varepsilon$ gives $u_0 + u_1 v_1 + \dots + u_r v_r = 0$, so, as u_i are F -independent, $v_k \notin F$ for some $k \leq r$. By the definition of $F = E_G$, $\tau v_k \neq v_k$ for some $\tau \in G$. Apply τ to equations $(**)$ to get

$$\tau \sigma u_0 + \tau \sigma u_1 \tau v_1 + \dots + \tau \sigma u_r \tau v_r = 0, \quad \sigma \in G.$$

Because $\tau \sigma$ runs through the entire group G as σ does, these equations, written in a different order, take the form

$$\sigma u_0 + \sigma u_1 \tau v_1 + \dots + \sigma u_r \tau v_r = 0, \quad \sigma \in G. \quad (***)$$

Now subtract $(***)$ from $(**)$ to get

$$\sigma u_1(v_1 - \tau v_1) + \dots + \sigma u_r(v_r - \tau v_r) = 0, \quad \sigma \in G.$$

As $v_k - \tau v_k \neq 0$, this gives a nontrivial solution to (*) with at most r nonzero values, contradicting the choice of r . \blacksquare

Example 1. Let $u = e^{2\pi i/5}$, $E = \mathbb{Q}(u)$, and $F = \mathbb{Q}$. If $G = \text{gal}(E : F)$, we showed in Example 4 §10.1 that $G = \langle \sigma \rangle \cong C_4$, where σ is defined by $\sigma(u) = u^2$. Thus, $[E : E_G] = 4$ by the Dedekind–Artin theorem. However, the minimal polynomial of u is $1 + x + x^2 + x^3 + x^4$, so $[E : \mathbb{Q}] = 4$. As $\mathbb{Q} \subseteq E_G \subseteq E$, this implies that $E_G = \mathbb{Q}$; that is, the only elements of E fixed by G are the elements of \mathbb{Q} .

Now consider $H = \langle \sigma^2 \rangle$ and compute $E_H = \{w \in E \mid \sigma^2(w) = w\}$. Note that $\{1, u, u^2, u^3\}$ is a \mathbb{Q} -basis of E and that $\sigma^2(u^k) = u^{4k}$ for each k (because $\sigma^2(u) = u^4$). If $w = a + bu + cu^2 + du^3$ is in E_H , this gives

$$w = \sigma^2(w) = a + bu^4 + cu^8 + du^{12} = a + b(-1 - u - u^2 - u^3) + cu^3 + du^2.$$

Then equating coefficients of the powers of u implies that $b = 0$ and $d = c$, so $w = a + c(u^2 + u^3)$. Thus, $[E_H : \mathbb{Q}] = 2$, and so $[E : E_H] = [E : \mathbb{Q}] / [E_H : \mathbb{Q}] = 2 = |H|$, as the Dedekind–Artin theorem asserts. \square

Galois Extensions

Fix a particular field extension $E \supseteq F$ and write $G = \text{gal}(E : F)$. A field K such that $E \supseteq K \supseteq F$ is called an **intermediate field** of the extension. The heart of Galois theory is the observation that these intermediate fields are intimately related to the subgroups of the Galois group G . Indeed, if K is an intermediate field, $\text{gal}(E : K)$ is a subgroup of G denoted, for convenience, by

$$K' = \text{gal}(E : K) = \{\sigma \in G \mid \sigma(u) = u \text{ for all } u \in K\} = \{\sigma \in G \mid \sigma \text{ fixes } K\}.$$

Conversely, for a subgroup H of G , the fixed field E_H of H in E is easily verified to be an intermediate field of the extension and is denoted for our present purposes as

$$H^\circ = E_H = \{u \in E \mid \sigma(u) = u \text{ for all } \sigma \in H\} = \{u \in E \mid u \text{ is fixed by } H\}.$$

The basic properties of these constructions are collected in Lemma 2.

Lemma 2. Let $E \supseteq F$ be fields and write $G = \text{gal}(E : F)$. Let K and K_1 be intermediate fields and let H and H_1 be subgroups of G . Then

- (1) If $K \subseteq K_1$, then $K' \supseteq K'_1$.
- (2) If $H \subseteq H_1$, then $H^\circ \supseteq H_1^\circ$.
- (3) $K \subseteq K'^\circ$ and $K'^\circ = \{u \in E \mid \text{if } \sigma \in G \text{ fixes } K \text{ then } \sigma \text{ fixes } u\}$.
- (4) $H \subseteq H'^\circ$ and $H'^\circ = \{\sigma \in G \mid \text{if } u \in E \text{ is fixed by } H, \text{ then } u \text{ is fixed by } \sigma\}$.
- (5) $K' = K'^\circ$.
- (6) $H^\circ = H'^\circ$.

Proof. (1) and (2) are immediate consequences of the definition, as are the descriptions of K'° and H'° in (3) and (4). These descriptions imply that $K \subseteq K'^\circ$ and $H \subseteq H'^\circ$, proving (3) and (4). Now $K' \supseteq K'^\circ$ and $H^\circ \supseteq H'^\circ$ by (1) and (2). But

$$K'^\circ = (K')^\circ \supseteq K' \text{ by (4)} \quad \text{and} \quad H'^\circ = (H^\circ)^\circ \supseteq H^\circ \text{ by (3)}.$$

This proves (5) and (6). \blacksquare

By virtue of these properties, the maps $K \mapsto K'$ and $H \mapsto H^\circ$ are called a **Galois connection**. The most interesting case is when these maps are mutually inverse bijections and we characterize the extensions for which this happens in a moment. However, we first need to say something about the most general case.

Following Irving Kaplansky,¹¹⁶ it is convenient to call H^o and K'^o the **closures** of H and K , respectively, and to call H and K **closed** if $H = H^o$ and $K = K'^o$, respectively. Thus, (5) and (6) of Lemma 2 assert that K' and H° are always closed, which leads to Lemma 3.

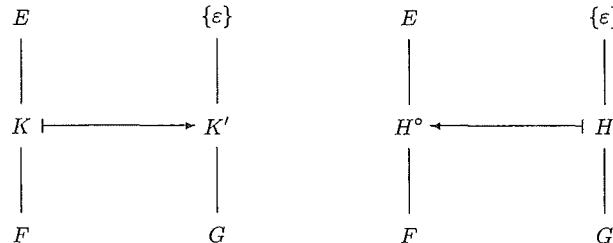
Lemma 3. *Let $E \supseteq F$ be fields and let $G = \text{gal}(E : F)$. Then*

$$K \mapsto K' \quad \text{and} \quad H \mapsto H^\circ$$

are mutually inverse, order reversing bijections between the set of closed intermediate fields K of the extension $E \supseteq F$ and the set of closed subgroups H of the Galois group G .

Proof. These maps are defined because K' and H° are closed, they are order reversing by (1) and (2) of Lemma 2, and they are mutually inverse bijections because $K'^o = K$ and $H^o = H$ whenever K and H are closed. ■

This result is slick, but it is not very useful unless we have a good idea about which intermediate fields and which subgroups are closed. To motivate the discussion, view the effect of the ' $'$ and ' \circ ' operations as shown in the diagram, where $E \supseteq F$ are fields and $G = \text{gal}(E : F)$.



Applying the operations at the tops and bottoms of these diagrams, one sees

$$\begin{aligned} E' &= \{\epsilon\} & \text{and} & \quad \{\epsilon\}^\circ = E, \\ F' &= G & \text{and} & \quad G^\circ \supseteq F. \end{aligned}$$

The anomaly $G^\circ \supseteq F$ begs for attention. It need not be equality: If $F = \mathbb{Q}$ and $E = \mathbb{Q}(\sqrt[3]{2})$, then $G = \{\epsilon\}$ by Example 3 §10.1, so $G^\circ = E \supsetneq F$. However, we do have some useful conditions when equality happens.

Lemma 4. *If $E \supseteq F$ are fields and $G = \text{gal}(E : F)$, the following are equivalent.*

- (1) $G^\circ = F$.
- (2) F is closed.
- (3) The only elements of E fixed by each $\sigma \in G$ are the elements of F .

¹¹⁶Kaplansky, I., *Fields and Rings*, 2nd ed., Chicago: University of Chicago Press, 1972.

Proof. As $F' = G$, we have $F'^\circ = G^\circ$, so (1) \Leftrightarrow (2). Finally, (1) \Leftrightarrow (3) follows because $G^\circ = E_G = \{u \in E \mid \sigma(u) = u \text{ for all } \sigma \in G\}$. ■

A field extension $E \supseteq F$ is called a **Galois extension** if the conditions in Lemma 4 are satisfied. Hence, $E \supseteq F$ is Galois if and only if

Given $u \in E$ with $u \notin F$, there exists $\sigma \in \text{gal}(E : F)$ such that $\sigma(u) \neq u$.

Example 2. $\mathbb{C} \supseteq \mathbb{R}$ is Galois because $\text{gal}(\mathbb{C} : \mathbb{R}) = \{\varepsilon, \gamma\}$, where γ is conjugation (Example 2 §10.1), and the only complex numbers fixed under conjugation are real.

Example 3. If $E = \mathbb{Q}(\sqrt[3]{2})$, then $E \supseteq \mathbb{Q}$ is a finite extension that is separable (as $\text{char } \mathbb{Q} = 0$) but is not Galois. Indeed, $G = \text{gal}(E : \mathbb{Q}) = \{\varepsilon\}$ by Example 3 §10.1, so every element of E is fixed by G .

Galois extensions have been defined very abstractly. Hence, Theorem 3 is fundamental because it characterizes them in terms of splitting fields and separability.

Theorem 3. The following conditions are equivalent for a finite field extension $E \supseteq F$ with Galois group $G = \text{gal}(E : F)$.

- (1) $E \supseteq F$ is a Galois extension.
- (2) Each irreducible polynomial in $F[x]$ with a root in E is separable and splits in $E[x]$.
- (3) E is the splitting field of some separable polynomial in $F[x]$.

In addition, $E \supseteq F$ is a separable extension.

Proof. (1) \Rightarrow (2). Let p be irreducible in $F[x]$ and let $p(u) = 0$, $u \in E$. Write $X = \{\tau(u) \mid \tau \in G\}$. Then X is a finite subset of E because G is a finite group by Theorem 2 §10.1 ($E \supseteq F$ is finite by hypothesis). So let $u = u_1, u_2, \dots, u_m$ denote the distinct elements of X and define f in $E[x]$ by

$$f = (x - u_1)(x - u_2) \cdots (x - u_m).$$

If $\sigma \in G$, then $\sigma u_1, \sigma u_2, \dots, \sigma u_m$ are distinct and they are elements of X because G is a group. Hence, they are the elements u_1, u_2, \dots, u_m in a different order. Since $f \mapsto f^\sigma$ is a ring isomorphism $F[x] \rightarrow F[x]$,¹¹⁷ this means that

$$f^\sigma = \prod_{i=1}^n (x - u_i)^\sigma = \prod_{i=1}^n (x - \sigma u_i) = \prod_{i=1}^n (x - u_i) = f.$$

It follows that σ fixes each coefficient of f . Hence, these coefficients lie in F by (1), and so $f \in F[x]$. But p is the minimal polynomial of u in F (being irreducible), so $f(u) = 0$ implies that p divides f in $F[x]$. Hence, p splits in E and is separable (because u_i are distinct). This proves (2).

(2) \Rightarrow (3). If $E = F$, there is nothing to prove. Otherwise, choose $u_1 \in E$, $u_1 \notin F$, and let p_1 be its minimal polynomial over F . Then p_1 is separable and splits over E by (2), so let $E_1 \subseteq E$ be a splitting field of p_1 . If $E_1 = E$, we are done; otherwise, choose $u_2 \in E$, $u_2 \notin E_1$, with minimal polynomial p_2 over F . Again, p_2 is separable and splits over E by (2). Write $f_2 = p_1 p_2$, so that f_2 is separable and splits over E . Let $E_2 \subseteq E$ be a splitting field of f_2 . If $E_2 = E$, we are done. This process must stop because $F \subset E_1 \subset E_2 \subset \cdots \subseteq E$ and $[E : F]$ is finite by hypothesis.

¹¹⁷If $f = a_0 + a_1 x + \cdots + a_n x^n$, then $f^\sigma = \sigma(a_0) + \sigma(a_1)x + \cdots + \sigma(a_n)x^n$.

(3) \Rightarrow (1) Consider $E \supseteq E_G \supseteq F$. Given (3), the corollary of Theorem 5 §10.1 shows that $[E : F] = |G|$. But $[E : E_G] = |G|$ by the Dedekind–Artin theorem, so $E_G = F$ and (1) follows because $G^\circ = E_G$ by definition.

Finally, $E \supseteq F$ is separable by the proof of (1) \Rightarrow (2). ■

Every algebraic extension of a field of characteristic 0 is separable, so we have

Corollary 1. *If $\text{char } F = 0$, the Galois finite extensions of F are precisely the splitting fields of polynomials in $F[x]$.*

Every finite Galois extension¹¹⁸ is separable by Theorem 3 (but see Example 3). However, the primitive element theorem (Theorem 6 §10.1) gives

Corollary 2. *Every finite Galois extension $E \supseteq F$ is simple; that is, $E = F(u)$ for some $u \in E$.*

The next corollary proves again the corollary to Theorem 5 §10.1.

Corollary 3. *If $E \supseteq F$ is a finite Galois extension and $G = \text{gal}(E : F)$, then $[E : F] = |G|$.*

Proof. The proof of (3) \Rightarrow (1) in Theorem 3 shows that $E_G = F$, so the Dedekind–Artin theorem applies. ■

Corollary 4. *If $E \supseteq K \supseteq F$ are fields and $E \supseteq F$ is a finite Galois extension, then $E \supseteq K$ is also a Galois extension.*

Proof. As $E \supseteq F$ is Galois, let E be the splitting field over F of the separable polynomial f in $F[x]$. Then $f \in K[x]$, E is a splitting field of f over K , and f is separable over K . Hence, $E \supseteq K$ is Galois by Theorem 3. ■

If $K \supseteq F$ is a finite Galois extension with intermediate field K , the easiest way to obtain elements in $\text{gal}(K : F)$ is often as the restriction to K of automorphisms in $\text{gal}(E : F)$ for some field $E \supseteq K$. Hence, Corollary 5 is useful because it places no condition on the extension $E \supseteq F$.

Corollary 5. *Let $E \supseteq K \supseteq F$ be fields, where $K \supseteq F$ is finite and Galois. If $\sigma \in \text{gal}(E : F)$, then $\sigma(K) = K$, so σ restricts to an automorphism in $\text{gal}(K : F)$.*

Proof. If $u \in K$, let p be its minimal polynomial over F . Given $\sigma \in \text{gal}(E : F)$, we have $p[\sigma(u)] = \sigma[p(u)] = \sigma(0) = 0$, so $\sigma(u) \in E$ is also a root of p . But p splits in K by Theorem 3, so $\sigma(u) \in K$. This proves that $\sigma(K) \subseteq K$. Similarly, $\sigma^{-1}(K) \subseteq K$, so $\sigma(K) = K$, as asserted. ■

Example 4. If $\mathbb{C} \supseteq K \supseteq \mathbb{Q}$, where K is the splitting field of a polynomial in $\mathbb{Q}[x]$, then $K \supseteq \mathbb{Q}$ is Galois (Corollary 1), so complex conjugation restricts to an automorphism in $\text{gal}(K : \mathbb{Q})$ by Corollary 5.

The Main Theorem

Until now, all our results have been valid for arbitrary subgroups of the Galois group. However, the normal subgroups play a special role, and the property in Corollary 5 is the analogue for intermediate fields of normality for subgroups. Again,

¹¹⁸Sometimes called a *normal extension*. However, the term “normal” is used in other ways.

our terminology follows Kaplansky. If $E \supseteq K \supseteq F$ are fields, the intermediate field K is called **stable** in the extension $E \supseteq F$ if

$$\begin{aligned}\sigma(K) &\subseteq K \text{ for all } \sigma \in \text{gal}(E : F), \text{ equivalently} \\ \sigma(K) &= K \text{ for all } \sigma \in \text{gal}(E : F).\end{aligned}$$

Clearly, both E and F are stable in $E \supseteq F$. And if $E \supseteq F$ is Galois, then every intermediate field K is stable in $E \supseteq F$ (Corollary 5).

Lemma 5. *Let $E \supseteq F$ be fields and let $G = \text{gal}(E : F)$. Then:*

- (1) *If H is a normal subgroup of G , then $H^\circ = E_H$ is stable in $E \supseteq F$.*
- (2) *If K is a stable intermediate field, then $K' = \text{gal}(E : K)$ is normal in G and $G/K' \cong \{\lambda \in \text{gal}(K : F) \mid \lambda \text{ extends to an automorphism of } E\}$.*

Proof. (1) Given $H \triangleleft G$, let $\sigma \in G$. We must show that $\sigma(H^\circ) \subseteq H^\circ$, that is, $\sigma(u) \in H^\circ$ for all $u \in H^\circ$. But if $\tau \in H$, then $\sigma^{-1}\tau\sigma \in H$, so $(\sigma^{-1}\tau\sigma)(u) = u$. Thus, $\tau[\sigma(u)] = \sigma(u)$ for all $\tau \in H$; that is, $\sigma(u) \in H^\circ$.

(2) If K is stable and $\sigma \in G$, then $\sigma(K) = K$, so the restriction $\sigma|_K : K \rightarrow K$ of σ to K is in $\text{gal}(K : F)$. Hence, define $\varphi : G \rightarrow \text{gal}(K : F)$ by $\varphi(\sigma) = \sigma|_K$ for all $\sigma \in G$. This is a group homomorphism and

$$\ker \varphi = \{\sigma \in G \mid \sigma(u) = u \text{ for all } u \in K\} = K'.$$

Finally, $\varphi(G) = \{\lambda \in \text{gal}(K : F) \mid \lambda \text{ extends to an automorphism of } E\}$. ■

Finally, we are ready to prove the most important theorem of this chapter. Recall that $|G : H|$ denotes the index of H in G where $H \subseteq G$ are finite groups.

Theorem 4. The Main Theorem of Galois Theory. *Let $E \supseteq F$ be a finite Galois extension with Galois group $G = \text{gal}(E : F)$, let K and K_1 denote intermediate fields of the extension $E \supseteq F$, and let H and H_1 denote subgroups of G . As before, write $K' = \text{gal}(E : K)$ and $H^\circ = E_H$.*

- (1) *All H and K are closed. The maps $K \mapsto K'$ and $H \mapsto H^\circ$ are mutually inverse, order-reversing bijections between the set of all intermediate fields K of the extension $E \supseteq F$ and the set of subgroups H of G .*
- (2) *If $K_1 \subseteq K$, then $[K : K_1] = |K'_1 : K'|$.*
- (3) *If $H_1 \subseteq H$, then $|H : H_1| = |H_1^\circ : H^\circ|$.*
- (4) *$E \supseteq K$ is a Galois extension.*
- (5) *$K \supseteq F$ is Galois \Leftrightarrow if K is stable in $E \supseteq F \Leftrightarrow K' \triangleleft G$.
In this case, $G/K' \cong \text{gal}(K : F)$.*

Proof. Observe first that (4) is Corollary 4 of Theorem 3.

(1) By Lemma 3, it suffices to show that all K and H are closed. We have $H \subseteq H^\circ$ by Lemma 2; to prove equality, note that $|H| = [E : H^\circ]$ by the Dedekind–Artin theorem. Replacing H by H° , we get $|H^\circ| = [E : H^{\circ\circ}] = [E : H^\circ] = |H|$. Hence, $H = H^\circ$ and H is closed. Turning to K , write $H_1 = \text{gal}(E : K)$. Then $E \supseteq K$ is Galois by (4). Hence, $K = \{u \in E \mid \sigma(u) = u \text{ for all } \sigma \in H_1\} = H_1^\circ$ and this is always closed by Lemma 2.

(2) By (4) $E \supseteq K$ is Galois, so $[E : K] = |\text{gal}(E : K)| = |K'|$ by Corollary 3 of Theorem 3. Similarly, $[E : K_1] = |K'_1|$. Hence,

$$[K : K_1] = \frac{[E : K_1]}{[E : K]} = \frac{|K'_1|}{|K'|} = |K'_1 : K'|.$$

(3) Write $H^\circ = K$ and $H_1^\circ = K_1$. Then $K \subseteq K_1$ and (1) gives $K' = H^{\circ\circ} = H$ and $K'_1 = H_1^{\circ\circ} = H_1$. Hence, (2) implies that

$$[H : H_1] = |K' : K'_1| = [K_1 : K] = [H_1^\circ : H^\circ].$$

(5) Since K is closed by (1), we have $K'^\circ = K$. Then Lemma 5 shows that K is stable if and only if $K' \triangleleft G$, and also gives $G/K' \cong \text{gal}(K : F)$. If $K \supseteq F$ is Galois, then K is stable by Corollary 5 of Theorem 3. Conversely, if K is stable, let $u \in K \setminus F$. As $E \supseteq F$ is Galois, σ in G exists such that $\sigma(u) \neq u$. But the restriction of σ to K is in $\text{gal}(K : F)$ because K is stable, and hence $K \supseteq F$ is Galois. ■

The main theorem has many uses because many properties of intermediate fields can be deduced from the analogous properties of the subgroups of the Galois group. To illustrate, we reprove an important property of finite fields (Theorem 5 §6.4).

Corollary. *If $E = GF(p^n)$, where p is a prime, then $E \supseteq \mathbb{Z}_p$ is Galois and the subfields of E are precisely the fields $GF(p^m)$, where $m|n$.*

Proof. E is the splitting field of $f = x^{p^n} - x$ over \mathbb{Z}_p (Theorem 4 §6.4) and $f' = -1$, so f has no repeated roots in E (Theorem 3 §6.4). Hence, $E \supseteq \mathbb{Z}_p$ is Galois. Next, $G = \text{gal}(E : \mathbb{Z}_p) \cong C_n$ by Example 6 §10.1. Thus, G has exactly one subgroup of order m for each divisor m of $|G| = [E : \mathbb{Z}_p] = n$, so the main theorem gives exactly one intermediate field K with $[E : K] = m$. It must be $GF(p^m)$. ■

The main theorem shows that if $E \supseteq F$ is a finite Galois extension, the lattice of intermediate fields has the same form as the (inverted) lattice of subgroups of $G = \text{gal}(E : F)$. Moreover, if H is a subgroup, then $H = K'$, where $K = H^\circ$. Hence, (5) of the main theorem translates to

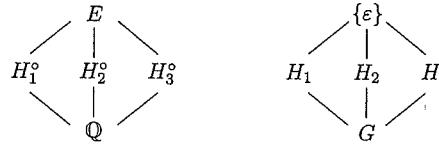
$$H \triangleleft G \quad \text{if and only if} \quad H^\circ \supseteq F \text{ is Galois.}$$

Examples 5 and 6 illustrate this.

Example 5. Let $E = \mathbb{Q}(u, v)$, where $u = \sqrt{2}$ and $v = \sqrt{3}$, and let $G = \text{gal}(E : \mathbb{Q})$. Show that $G = \langle \sigma, \tau \rangle \cong C_2 \times C_2$, where σ and τ are defined by $\sigma(u) = u$, $\sigma(v) = -v$, and $\tau(u) = -u$, $\tau(v) = v$. Hence, find all the intermediate fields in $E \supseteq \mathbb{Q}$.

Solution. The minimal polynomial of v is $x^2 - 3$, and the other root is $-v$. Hence, there is a \mathbb{Q} -automorphism $\sigma_0 : \mathbb{Q}(v) \rightarrow \mathbb{Q}(-v)$ satisfying $\sigma_0(v) = -v$. This mapping extends to an automorphism σ of $E = \mathbb{Q}(v)(u) = \mathbb{Q}(-v)(u)$ that satisfies $\sigma(u) = u$. This creates σ , and we construct τ in the same way. Now $|G| \leq 4$ because $\lambda(u) = \pm u$ and $\lambda(v) = \pm v$ for all $\lambda \in G$. Because $o(\sigma) = 2 = o(\tau)$ and $\sigma\tau = \tau\sigma$, it follows that $G = \langle \sigma, \tau \rangle \cong C_2 \times C_2$.

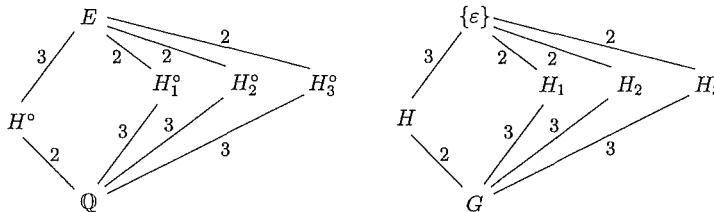
The subgroups of G are $\{\varepsilon\}$, G , $H_1 = \langle \sigma \rangle$, $H_2 = \langle \tau \rangle$, and $H_3 = \langle \sigma\tau \rangle$. The subgroup lattice (inverted) is shown in the right-hand diagram. Hence, the lattice of intermediate fields is as shown in the left-hand diagram.



Now $E \supseteq \mathbb{Q}$ is Galois— E is the splitting field of $(x^2 - 2)(x^2 - 3)$, so the main theorem ensures that all intermediate fields can be obtained in this way. Also, $H_i \triangleleft G$ for each i guarantees that $H_i^o \supseteq \mathbb{Q}$ is Galois.

Finally, the main theorem is useful in the actual computation of the intermediate fields. Clearly, $u \in H_1^o$, so $\mathbb{Q}(u) \subseteq H_1^o$. But $[\mathbb{Q}(u) : \mathbb{Q}] = 2$ because $x^2 - 2$ is the minimal polynomial of u , and so $[H_1^o : \mathbb{Q}] = [H_1^o : G^o] = |G : H_1| = 2$ by the main theorem. Hence, $H_1^o = \mathbb{Q}(u)$, and similar arguments give $H_2^o = \mathbb{Q}(v)$ and $H_3^o = \mathbb{Q}(uv)$ because $\sigma\tau(uv) = uv$. \square

Example 6. Let E be the splitting field of $x^3 - 2$ over \mathbb{Q} . Thus, $E \supseteq \mathbb{Q}$ is a Galois extension by Theorem 3 because $\text{char } \mathbb{Q} = 0$. In Example 7 §10.1, we showed that $G = \text{gal}(E : \mathbb{Q}) \cong D_3$. In fact, write $u = \sqrt[3]{2}$ and $w = e^{2\pi i/3}$. Then $G = \langle \sigma, \tau \rangle$, where $\sigma(\sigma) = 3$, $\sigma(\tau) = 2$, and $\sigma\tau\sigma = \tau$ and σ and τ are defined by $\sigma(u) = uw$ and $\sigma(w) = w$, whereas $\tau(u) = u$ and $\tau(w) = w^2$. Thus, the subgroups of G are $\{\epsilon\}$, G , $H = \langle \sigma \rangle$, $H_1 = \langle \tau \rangle$, $H_2 = \langle \tau\sigma \rangle$, and $H_3 = \langle \tau\sigma^2 \rangle$. The (inverted) subgroup lattice is shown in the right diagram along with the index of each group extension. The left diagram gives the lattice of intermediate fields of the extension $E \supseteq \mathbb{Q}$ along with all the dimensions. The main theorem guarantees that the dimensions and indices correspond, as indicated in the figure. Moreover, the fact that H is normal in G but H_1 , H_2 , and H_3 are not shows that $H^o \supseteq \mathbb{Q}$ is a Galois extension, whereas $H_1^o \supseteq \mathbb{Q}$, $H_2^o \supseteq \mathbb{Q}$, and $H_3^o \supseteq \mathbb{Q}$ are not.



Again the main theorem is useful in the computation of the intermediate fields. We have $\sigma(w) = w$ so, as $H = \langle \sigma \rangle$, $\mathbb{Q}(w) \subseteq H^o$. But $[\mathbb{Q}(w) : \mathbb{Q}] = 2$ because the minimal polynomial of w is $x^2 + x + 1$, and $[H^o : \mathbb{Q}] = [H^o : G^o] = |G : H| = 2$ by the main theorem. Hence, $H^o = \mathbb{Q}(w)$. Similarly, $\tau(u) = u$ implies that $H_1^o = \mathbb{Q}(u)$. To find primitive elements for H_2^o and H_3^o , we must find elements of $E \setminus \mathbb{Q}$ fixed by $\tau\sigma$ and $\tau\sigma^2$, respectively. Now each of these automorphisms must permute the set $\{u, uw, uw^2\}$ of roots of $x^3 - 2$ and because these maps have order 2 in G , each must fix one of these roots. A routine check reveals that $\tau\sigma(uw) = uw$ and $\tau\sigma^2(uw^2) = uw^2$, so $H_2^o = \mathbb{Q}(uw)$ and $H_3^o = \mathbb{Q}(uw^2)$. \square

The main theorem is not the end of the Galois theory, but rather the beginning. The study of abelian Galois groups leads to class field theory, an active research area with applications to algebraic number theory. If fields are replaced

by commutative rings¹¹⁹ or by division rings,¹²⁰ much of the theory still applies, suitably modified, and a version of the main theorem holds in each case. Ritt and Kolchin¹²¹ developed a differential Galois theory in which differential equations replace polynomial equations and in which a type of main theorem is proven. Other Galois-type theories also exist; the idea of a Galois correspondence occurs frequently and often gives important information about the objects that correspond.

Évariste Galois (1811–1832) Galois was born near Paris to well-educated parents and after tutoring by his mother, entered school at the age of 12. His routine school work was mediocre, but he discovered Legendre's *Éléments de Géométrie*, which captivated him; it is said he read it like a novel and mastered it in one reading. He then went on to works of Lagrange and Abel and at the age of 15, was reading professional-level material and beginning to make discoveries of his own. Unfortunately, his work was not systematic, with much of the calculation done mentally and only the results written down. He tried twice to enter the École Polytechnique but was rejected because of his lack of systematic preparation. This rejection was a great loss for mathematics because the École, which had produced many great mathematicians, may have been able to recognize his genius and provide the environment he needed.

Nonetheless, Galois continued to make fundamental discoveries about polynomial equations and, in 1829, submitted some of his results to the Académie des Sciences. The referee was Cauchy, who was certainly competent to understand it, but Cauchy lost the manuscript and it was never seen again! Undaunted, Galois submitted his work in the 1830 competition for the Académie's grand prize in mathematics. The article should have won this highest honor for its author, but the secretary, Fourier, took the manuscript home and, incredibly, died before reading it. The manuscript was lost. Finally, Galois sent a second memoir to the Académie. This time Poisson reviewed it and declared it to be "incomprehensible."

Whether because of bitterness over these events or because of his father's republican sympathies, Galois reacted by blaming the Bourbon regime and joining the National Guard, a republican organization. It was a time of great political unrest in France, and, as a result, he was in and out of prison, regaining his freedom in 1832. At this time, he became involved with a girl. The details of this liaison are obscure but one thing is certain: He was challenged to a duel and felt honor bound to go through with it. He had a sense of foreboding about the duel and wrote, "I die the victim of an infamous coquette. It is in a miserable brawl that my life is extinguished. Oh! why die for so trivial a thing...." The night before the duel, he wrote a letter to a friend outlining his discoveries. It is a tragic, poignant document with comments such as "I have no time" scribbled in the margins, and it ends by asking that Jacobi or Gauss give their opinion "not as to the truth, but as to the importance of these theorems." Hermann Weyl, one of the greatest mathematicians of the twentieth century, has written that "...this letter, if judged by the novelty and profundity of the ideas it contains, is perhaps the most substantial piece of writing in the whole literature of mankind."

¹¹⁹Chase, S.U., Harrison, D.K., and Rosenberg, A., *Galois Theory and Cohomology of Commutative Rings*, American Mathematical Society Memoir 52, Providence, RI: American Mathematical Society, 1965.

¹²⁰Jacobson, N., *Structure of Rings*, Colloquium Publications XXXVII, Providence, RI: American Mathematical Society, 1964.

¹²¹See Kaplansky, I., *An Introduction to Differential Algebra*, Paris: Hermann, 1957.

The duel was with pistols at 25 paces. Galois was hit in the stomach and lay where he fell until a passing peasant took him to a hospital. He died the next day, May 31, 1832, at the age of 20, and was buried in the common ditch at the cemetery of Montparnasse.

Exercises 10.2

1. In each case, show that $E \supseteq F$ is Galois, find the lattice of intermediate fields, and find a primitive element for each intermediate field.
 - (a) $E = \mathbb{Q}(u)$, where $u = e^{2\pi i/5}$, $F = \mathbb{Q}$.
 - (b) $E = \mathbb{Q}(u)$, where $u = e^{2\pi i/7}$, $F = \mathbb{Q}$.
 - (c) $E = \mathbb{Q}(i, \sqrt{3})$, $F = \mathbb{Q}$.
 - (d) $E = \mathbb{Z}_2(u)$, u a root of $x^4 + x + 1$, $F = \mathbb{Z}_2$.
 - (e) $E = \mathbb{Q}(\sqrt[4]{2}, i)$, $F = \mathbb{Q}$. [Hint: Exercise 13 §10.1.]
2. In each case, describe all possible intermediate field lattices for a finite Galois extension $E \supseteq F$.
 - (a) $|\text{gal}(E : F)| = p^2$, where p is a prime. [Hint: Theorem 7 §8.2.]
 - (b) $|\text{gal}(E : F)| = 2p$, where p is a prime. [Hint: Theorem 3 §2.6.]
3. If $E = GF(p^n)$, use the Dedekind–Artin theorem to show that $E \supseteq \mathbb{Z}_p$ is a Galois extension and display the lattice of subfields of $GF(p^{12})$ in terms of the Frobenius automorphism of E . [Hint: Example 6 §10.1 and Theorem 3.]
4. (a) If $H = \langle X \rangle$, $X \subseteq \text{gal}(E : F)$, show that $H^\circ = \{u \in E \mid \sigma(u) = u \text{ for all } \sigma \in X\}$.

(b) If X is finite and $K = \{u \in E \mid \sigma(u) = u \text{ for all } \sigma \in X\}$, show that K is an intermediate field and that $[E : K] \geq |X|$.
5. Let $E = F(t)$ be the field of rational forms over a field. In each case, compute $K = E_G$ and find the minimal polynomial $m \in K[x]$ of t over K .
 - (a) $G = \langle \sigma \rangle$, where σ is that F -automorphism of E given by $\sigma(t) = -t$.
 - (b) $G = \langle \sigma \rangle$, where σ is that F -automorphism of E given by $\sigma(t) = 1 - t$.
6. Show that a finite Galois extension has a finite number of intermediate fields.
7. Let $E \supseteq K \supseteq F$ be fields. If $E \supseteq F$ is finite and Galois and if $\text{gal}(E : F)$ is abelian, show that $K \supseteq F$ is Galois.
8. Let $E \supseteq F$ be finite and Galois, where $\text{gal}(E : F)$ is cyclic. If k divides $[E : F]$, show that there is exactly one intermediate field K such that $[E : K] = k$.
9. Let $E \supseteq F$ be fields with $G = \text{gal}(E : F)$ and consider the Galois connection.
 - (a) Show that $H \mapsto H^\circ$ is onto if and only if every intermediate field is closed.
 - (b) Show that $K \mapsto K'$ is onto if and only if every subgroup of G is closed.
10. Let $E \supseteq F$ be fields with $G = \text{gal}(E : F)$. If $H \subseteq G$ is a subgroup and H° is finite, show that H is closed.
11. If $E \supseteq K \supseteq F$ are fields, show that $E \supseteq K$ is Galois if and only if K is closed as an intermediate field of $E \supseteq F$.
12. If $E \supseteq F$ is finite and $G = \text{gal}(E : F)$, show that $E \supseteq F$ is Galois if and only if $|G| = [E : F]$.
13. If $E \supseteq F$ is a finite Galois extension with $\text{gal}(E : F) \cong A_4$, show that there is no intermediate field K with $[E : K] = 6$. [Hint: Exercise 34 §2.6.]
14. Let $E \supseteq F$ be a finite Galois extension, write $G = \text{gal}(E : F)$, and consider the intermediate field $K = \{u \in E \mid \sigma\tau(u) = \tau\sigma(u) \text{ for all } \sigma, \tau \in G\}$. Show that $K \supseteq F$ is a Galois extension with abelian Galois group.

15. Let $E \supseteq F$ be a finite Galois extension. If K and L are intermediate fields, let $K \vee L$ denote the intersection of all intermediate fields containing K and L . The field $K \vee L$ is called the **compositum** of K and L .
- Show that $(K \vee L)' = K' \cap L'$.
 - Describe the group $(K \cap L)'$ in terms of K' and L' .
16. An extension $E \supseteq F$ is called **abelian** (respectively, **cyclic**) if it is finite, Galois, and the Galois group $G = \text{gal}(E : F)$ is abelian (respectively, cyclic). If $E \supseteq K \supseteq F$, where $E \supseteq F$ is abelian (respectively, cyclic), show that both $E \supseteq K$ and $K \supseteq F$ are abelian (respectively, cyclic).
17. Let K and K_1 be intermediate fields in a finite Galois extension $E \supseteq F$. Show that K' and K'_1 are conjugate subgroups of $G = \text{gal}(E : F)$ if and only if $K = \sigma(K_1)$ for some $\sigma \in G$. (K and K_1 are called **conjugate** intermediate fields in this case.)
18. Let $E \supseteq F$ be fields with $G = \text{gal}(E : F)$. If K is an intermediate field and $K \supseteq F$ is a finite Galois extension, show that $\sigma(K) = K$ for all $\sigma \in G$. [Hint: Theorem 3.]
19. Let $f \in F[x]$, let $E \supseteq F$ be a splitting field of f over F , and let $G = \text{gal}(E : F)$.
 - Show that G can be embedded in S_m , where f has m distinct roots in E . [Hint: Theorem 3 §10.1.]
 - If f is separable, show that $[E : F]$ divides $n!$ [Compare with Theorem 2 §6.3.]
20. Let $E \supseteq F$ be a finite Galois extension with Galois group $G = \text{gal}(E : F)$. If $u \in E$, define the **norm** $N(u) = N_{E/F}(u)$ and the **trace** $T(u) = T_{E/F}(u)$ by
- $$N(u) = \prod_{\sigma \in G} \sigma(u) \quad \text{and} \quad T(u) = \sum_{\sigma \in G} \sigma(u).$$
- Show that $N(u)$ and $T(u)$ are in F . [Hint: $G^\circ = F$.]
 - Show that $N(uv) = N(u)N(v)$ and $T(u+v) = T(u) + T(v)$ for all $u, v \in E$.
 - Let $K = F(u)$ and let $p = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ be the minimal polynomial of u over F . If $K \supseteq F$ is Galois, show that $N_{K/F}(u) = (-1)^n a_0$ and $T_{K/F}(u) = -a_{n-1}$.
21. Let $E \supseteq F$ be a finite Galois extension with Galois group G . If $u \in E$, define $f = \prod_{\sigma \in G} (x - \sigma(u))$. Show that $f \in F[x]$ and is a power of the minimal polynomial of u over F .

10.3 INSOLVABILITY OF POLYNOMIALS

Possibly the best known result in algebra is the formula for the roots u_1 and u_2 of the equation $x^2 + bx + c = 0$. By *completing the square*, we write it in the form

$$(x + \frac{1}{2}b)^2 = \frac{1}{4}(b^2 - 4c).$$

Hence, we obtain the roots from

$$u_1 = \frac{1}{2}(-b + \sqrt{b^2 - 4c}) \quad \text{and} \quad u_2 = \frac{1}{2}(-b - \sqrt{b^2 - 4c}).$$

This is called the *quadratic formula* and was known in antiquity. The expression $\Delta = b^2 - 4c$ is called the *discriminant* of the quadratic $x^2 + bx + c$.

It was not until the sixteenth century that such a formula for the cubic was found. Given $y^3 + ry^2 + sy + t$, the substitution $y = x - \frac{1}{3}r$ gives $y^3 + ry^2 + sy + t = x^3 + bx + c$ for appropriate b and c . Hence, we need to find only formulas for the roots of cubic equations of the form

$$x^3 + bx + c = 0.$$

In this case, if the roots are u_1, u_2 , and u_3 , the cubic factors as

$$x^3 + bx + c = (x - u_1)(x - u_2)(x - u_3),$$

so the roots are related to the coefficients as follows:

$$\begin{aligned} u_1 + u_2 + u_3 &= 0, \\ u_1 u_2 + u_1 u_3 + u_2 u_3 &= b, \\ u_1 u_2 u_3 &= -c. \end{aligned}$$

Now let $w = e^{2\pi i/3}$ be a cube root of unity, so that $w^3 = 1$ and $1 + w + w^2 = 0$. We look for formulas for the roots of the form

$$u_1 = p + q, \quad u_2 = wp + w^2q, \quad \text{and} \quad u_3 = w^2p + wq. \quad (*)$$

Here, p and q are to be determined. Then the condition $u_1 + u_2 + u_3 = 0$ is automatically satisfied because $1 + w + w^2 = 0$, and the other two requirements reduce to $pq = -b/3$ and $p^3 + q^3 = -c$, respectively. These equations imply that p^3 and q^3 both satisfy the quadratic equation $x^2 + cx - b^3/27 = 0$. The quadratic formula then gives

$$p = \left[\frac{1}{2} \left(-c + \sqrt{c^2 + \frac{4}{27}b^3} \right) \right]^{1/3} \quad \text{and} \quad q = \left[\frac{1}{2} \left(-c - \sqrt{c^2 + \frac{4}{27}b^3} \right) \right]^{1/3}.$$

If we choose the cube roots so that $pq = -b/3$, then $(*)$ gives the three roots of $x^3 + bx + c = 0$. This expression is called the **cubic formula** and was first discussed by Scipione del Ferro (ca. 1465–1526). Incidentally, the quantity $\Delta = -4b^3 - 27c^2$ is called the **discriminant** of $x^3 + bx + c$, so the quantities p and q are given by

$$\sqrt[3]{\frac{1}{2} \left(-c \pm \sqrt{-\frac{1}{27}\Delta} \right)}.$$

Niccolò Tartaglia later rediscovered the cubic formula, and Girolamo Cardano published it in 1545 in his book *Ars Magna*. The book also contained Lodovico Ferrari's method for solving quartic equations, which led to many attempts in the seventeenth and eighteenth centuries to find a formula for the solution of quintic equations. Both Euler and Lagrange tried it and failed, although Lagrange succeeded in unifying the lower degree methods. In 1824, Abel gave the first conclusive proof that no such formula exists; the proof we give is due to Galois.

Clearly, the quadratic and cubic formulas are valid over any field F satisfying $\text{char } F \neq 2, 3$, and the roots can be found in an extension field of F obtained by adjoining square and cube roots of elements of F . If $E \supseteq F$ are fields, E is called a **radical extension** of F if a chain

$$E = E_0 \supseteq E_1 \supseteq E_2 \supseteq \cdots \supseteq E_n = F$$

of intermediate fields E_i exists such that

$$E_i = E_{i+1}(u_i), \quad \text{where } u_i^{n_i} \in E_{i+1} \text{ for some } n_i \geq 1.$$

A polynomial in $F[x]$ is called **solvable** over F if all its roots lie in some radical extension of F , equivalently if some radical extension of F contains a splitting field. Note that every radical extension is finite and every finite field is a radical extension of any subfield.

Thus, a polynomial f in $F[x]$ is solvable if and only if we can find the roots of f (in some splitting field) by using only operations of the field F and adjoining n th roots. Clearly, these operations yield the roots of quadratic and cubic polynomials (by the preceding formulas), so all quadratics and cubics are solvable (provided $\text{char } F \neq 2, 3$). This statement also holds for quartics¹²² but fails for quintics: There is a polynomial of degree 5 in $\mathbb{Q}[x]$ that is not solvable.

Let f be a polynomial in $F[x]$, where F is a field. If $E \supseteq F$ is any splitting field of f over F , the group $\text{gal}(E : F)$ is called the **Galois group of the polynomial** f . Note that the definition of the Galois group of a polynomial does not depend on which splitting field is used. In fact, if $\bar{E} \supseteq F$ is another splitting field of f , there is an F -isomorphism $E \rightarrow \bar{E}$, which implies that $\text{gal}(E : F) \cong \text{gal}(\bar{E} : F)$.

Galois' idea was to characterize solvable polynomials by a property of their Galois groups. Recall that a group G is called *solvable* if there is a chain of subgroups

$$G = G_0 \supseteq G_1 \cdots \supseteq G_n = \{1\}$$

such that $G_{i+1} \triangleleft G_i$ and G_i/G_{i+1} is abelian for each i . Clearly, every abelian group is solvable, and we discussed these groups at length in Section 9.2. The result we need is Theorem 4 §9.2, which we restate as Lemma 1 for reference.

Lemma 1. *If G is a group and K is a normal subgroup of G , then G is solvable if and only if both K and G/K are solvable.*

Note that the only solvable simple groups are abelian. Hence, the symmetric group S_n is not solvable if $n \geq 5$, because otherwise its normal subgroup A_n would be solvable by Lemma 1, contrary to the fact (Theorem 8 §2.8) that it is simple and nonabelian.

Now we can give Galois' approach to the insolvability of the general quintic. The key result is Galois' criterion.

Galois Criterion. *Let F be a field of characteristic 0. Then a polynomial in $F[x]$ is solvable over F if and only if its Galois group is a solvable group.¹²³*

Thus, Galois simply produced a polynomial of degree 5 in $\mathbb{Q}[x]$ whose Galois group is S_5 (and hence not solvable). Because this polynomial is not solvable, some root cannot be expressed using only rational operations and the extraction of n th roots. Clearly, only half of Galois' criterion is needed: Solvable polynomials have solvable Galois groups; this is, Theorem 2 (we do not prove the converse¹²⁴). Here is an example of a polynomial that is not solvable.

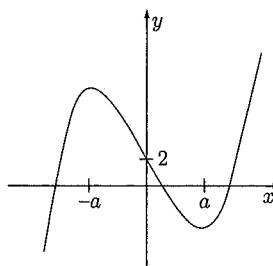
Example 1. Let $p = x^5 - 6x + 2$ in $\mathbb{Q}[x]$. Show that the Galois group of p is S_5 and hence that p is not solvable.

¹²²See, for example, Ehrlich, G., *Fundamental Concepts of Abstract Algebra*, Boston: PWS-KENT, 1991, p. 327.

¹²³This criterion is the source of the term *solvable group*.

¹²⁴See Rotman, J., *Galois Theory*, Berlin: Springer, 1990, p. 55.

Solution. We let $E \supseteq \mathbb{Q}$ be a splitting field of p so that $G = \text{gal}(E : \mathbb{Q})$ is the Galois group of p . We show that $G \cong S_5$ by identifying G as a group of permutations of the set X of roots of p in \mathbb{C} (see Theorem 3 §10.1). Now $p' = 5x^4 - 6$, which has real roots $\pm a$, $a = \sqrt[4]{6/5}$. We can easily verify that $p(a) < 0$ and $p(-a) > 0$, so the graph of p is as shown in the figure. In particular, p has three distinct real roots and two (conjugate) nonreal roots.¹²⁵ Hence, complex conjugation induces the transposition in G that exchanges the two nonreal roots (by Example 4 §10.2). However, p is irreducible by the Eisenstein criterion and so is the minimal polynomial of any root u in E . Hence, $[\mathbb{Q}(u) : \mathbb{Q}] = \deg p = 5$. But $|G| = [E : \mathbb{Q}]$ because $E \supseteq \mathbb{Q}$ is a Galois extension, so 5 divides $|G|$. Thus, G contains an element of order 5 by Cauchy's theorem (Theorem 4 §8.2). The only elements of order 5 in S_5 are the 5-cycles, which shows that G contains a 5-cycle and a 2-cycle (conjugation). Finally, this in turn implies that $G = S_5$ by Lemma 2 below, so $G \cong S_5$, as required. \square



The proof of the next theorem requires two lemmas that are of independent interest. The first involves the symmetric group S_p , where p is a prime.

Lemma 2. *If p is a prime, S_p is generated by any p -cycle and any 2-cycle.*

Proof. Choose the notation so that $\sigma = (1 \ 2 \ \cdots \ p)$ and $\tau = (1 \ k)$ are the given cycles. Now $\sigma^{k-1}(1) = k$ and σ^{k-1} is a p -cycle (as p is prime), so we may assume that $\sigma = (1 \ 2 \ \cdots \ p)$ and $\tau = (1 \ 2)$. Hence $(k+1 \ k+2) = \sigma^k \tau \sigma^{-k}$ for each k (by Lemma 3 §2.8). Because $(12), (13), \dots, (1 \ p)$ generate S_p and because we have $(1 \ a+1) = (1 \ a)(a \ a+1)(1 \ a)$, the proof is complete. \blacksquare

Lemma 3. *If G is a cyclic group of order n , then $\text{aut } G \cong \mathbb{Z}_n^*$.*

Proof. Since $G \cong \mathbb{Z}_n$, we show that $\text{aut } \mathbb{Z}_n \cong \mathbb{Z}_n^*$. If $m \in \mathbb{Z}_n^*$, define $\sigma_m : \mathbb{Z}_n \rightarrow \mathbb{Z}_n$ by $\sigma_m(k) = mk$. This is an automorphism of \mathbb{Z}_n because m is a unit in the ring \mathbb{Z}_n , so we have a mapping

$$\theta : \mathbb{Z}_n^* \rightarrow \text{aut } \mathbb{Z}_n \quad \text{given by} \quad \theta(m) = \sigma_m.$$

This is a homomorphism because $\sigma_{mm'} = \sigma_m \sigma_{m'}$, and it is one-to-one because $\sigma_m = \sigma_{m'}$ implies $m = \sigma_m(1) = \sigma_{m'}(1) = m'$. Finally, given $\sigma \in \text{aut } \mathbb{Z}_n$, write $m = \sigma(1)$. Then m is a generator of \mathbb{Z}_n and so $\gcd(m, n) = 1$. Thus $m \in \mathbb{Z}_n^*$, and $\sigma = \sigma_m$ because $\sigma(k) = \sigma(1)k = mk = \sigma_m(k)$ for all $k \in \mathbb{Z}_n$. Hence, $\sigma = \theta(m)$, so θ is an isomorphism. \blacksquare

As in the derivation of the cubic formula, the n th roots of unity (that is, the roots of $x^n - 1$) play an important role in the proof of the Galois criterion. If F is any field, a root of unity w in some extension field E is called a primitive n th root of unity over F if $|w| = n$ in E^* . Clearly, $w = e^{2\pi i/n}$ is an example in \mathbb{C} . In general, such an element w exists in an extension field of F if and only if $p = \text{char } F$ does not

¹²⁵The tacit assumption is that p splits in \mathbb{C} , which the fundamental theorem of algebra guarantees.

divide n (this is intended to include the case $\text{char } F = 0$). Indeed, if $n = pd$, then $x^n - 1 = (x^d - 1)^p$, so no primitive n th root of unity can exist over F . Conversely, if p does not divide n , then $x^n - 1$ and its derivative nx^{n-1} are relatively prime in $F[x]$, so $x^n - 1$ has n distinct roots in any splitting field $E \supseteq F$ (by Theorem 3 §6.4). These roots form a subgroup of E^* of order n that is cyclic by Theorem 7 §6.4. A generator of this group clearly is a primitive n^{th} root of unity, which proves the first statement in Theorem 1.

Theorem 1. *If F is a field and $n \geq 1$ is an integer, a primitive n th root of unity w over F exists if and only if $\text{char } F$ does not divide n . In this case,*

- (1) *$F(w)$ is the splitting field of $x^n - 1$ over F .*
- (2) *$F(w) \supseteq F$ is a finite Galois extension and $\text{gal}(F(w) : F)$ is isomorphic to a subgroup of \mathbb{Z}_n^* .*

Proof. Here, $F(w)$ is the splitting field of $x^n - 1$ over F because (as w is primitive) the roots are $1, w, \dots, w^{n-1}$, which all lie in $F(w)$. Moreover, these roots are distinct, so $x^n - 1$ is separable over F . Hence, $F(w) \supseteq F$ is Galois by Theorem 3 §10.2. Finally, if $\sigma \in \text{gal}[F(w) : F]$, then σ induces an automorphism σ_0 of $\langle w \rangle = \{1, w, \dots, w^{n-1}\}$ by restriction, and the map $\sigma \mapsto \sigma_0$ is a one-to-one group homomorphism. But $\text{aut}(\langle w \rangle) \cong \mathbb{Z}_n^*$ by Lemma 3, which completes the proof. ■

From the definition of a radical extension, any discussion of Galois' criterion clearly involves extensions $F(u) \supseteq F$, where $u^n \in F$. Lemma 4 will be needed.

Lemma 4. *Let F be a field containing a primitive n th root of unity and consider an extension $F(u)$, where $u^n \in F$. Then $F(u) \supseteq F$ is a Galois extension and $\text{gal}[F(u) : F]$ is abelian.*

Proof. Let $w \in F$ be a primitive n th root of unity and write $u^n = a \in F$. Then $x^n - a$ has roots u, uw, \dots, uw^{n-1} in $F(u)$, distinct because w is primitive. Hence, $x^n - a$ is separable over F , and $F(u)$ is the splitting field. Then $F(u) \supseteq F$ is Galois by Theorem 3 §10.2. Finally, if σ and τ are in $\text{gal}[F(u) : F]$, then $\sigma(u)$ and $\tau(u)$ are roots of $x^n - a$, say, $\sigma(u) = uw^i$ and $\tau(u) = uw^j$. Because $\sigma(w) = w = \tau(w)$, this gives $\tau\sigma(u) = uw^{i+j} = \sigma\tau(u)$. Thus, $\sigma\tau = \tau\sigma$, so $\text{gal}[F(u) : F]$ is abelian. ■

With this result, we can prove the half of Galois' criterion needed in Example 1.

Theorem 2. (Galois) *Let $E \supseteq F$ be a radical Galois extension, where $\text{char } F = 0$. Then $\text{gal}(E : F)$ is a solvable group.*

Proof. It suffices to find a field $K \supseteq E \supseteq F$, where $K \supseteq F$, is Galois and $\text{gal}(K : F)$ is solvable, because then $\text{gal}(E : F)$ is an image of $\text{gal}(K : F)$ by the main theorem. This uses the hypothesis that $E \supseteq F$ is Galois; we use the assumption that $E \supseteq F$ is radical to construct K . Let

$$E = E_0 \supseteq E_1 \supseteq \dots \supseteq E_r = F,$$

where $E_i = E_{i+1}(u_i)$ and $u_i^{n_i} \in E_{i+1}$ for $i = 0, 1, \dots, r-1$. Write $n = n_0 n_1 \cdots n_{r-1}$ and (as $\text{char } F = 0$) let w be a primitive n th root of unity over F . Define $K_i = E_i(w)$ for $0 \leq i \leq r$ and write $K = K_0 = E(w)$. Then K is the splitting field of $x^n - 1$ over E , and E is the splitting field over F of some f in $F[x]$ by Theorem 3 §10.2. Hence, K is the splitting field of $f(x)(x^n - 1)$ over F . Thus, $K \supseteq F$ is Galois (because

$\text{char } F = 0$), and it remains to show that $\text{gal}(K : F)$ is a solvable group. Write $K_{r+1} = F$ and consider the chain of fields

$$K = K_0 \supseteq K_1 \supseteq \cdots \supseteq K_r \supseteq K_{r+1} = F. \quad (**)$$

Claim 1. $K_i \supseteq K_{i+1}$ is Galois and $\text{gal}(K_i : K_{i+1})$ is abelian for each $i = 0, 1, 2, \dots, r$.

Proof. If $i = r$, the claim follows from Theorem 1 because $K_r = F(w) = K_{r+1}(w)$. If $i < r$, the claim follows from Lemma 4 because $K_i = E_{i+1}(u_i, w) = K_{i+1}(u_i)$; $u_i^{n_i} \in E_{i+1} \subseteq K_{i+1}$, and K_{i+1} contains a primitive n_i th root (namely, w^{n/n_i}). This proves the claim.

Now $(**)$ gives rise to the chain of Galois groups:

$$\{\varepsilon\} = \text{gal}(K : K_0) \subseteq \text{gal}(K : K_1) \subseteq \cdots \subseteq \text{gal}(K : K_r) \subseteq \text{gal}(K : K_{r+1}).$$

Because $K_{r+1} = F$, the proof is complete if we can show that $\text{gal}(K : K_i)$ is a solvable group for each $i = 0, 1, 2, \dots, r+1$. This is clear if $i = 0$, so assume inductively that $\text{gal}(K : K_i)$ is solvable and consider $K \supseteq K_i \supseteq K_{i+1}$. Then the extension $K \supseteq K_{i+1}$ is Galois by the main theorem (applied to $K \supseteq K_{i+1} \supseteq F$) and $K_i \supseteq K_{i+1}$ is Galois by the claim. Hence, the main theorem shows that $\text{gal}(K : K_i)$ is a normal subgroup of $\text{gal}(K : K_{i+1})$ and that the factor is isomorphic to $\text{gal}(K_i : K_{i+1})$ and so is abelian. Then $\text{gal}(K : K_{i+1})$ is solvable by Lemma 1, and the proof is complete. \blacksquare

Theorem 2 (together with Example 1) settles the question of solvability of polynomials of degree 5 in the negative, but it leaves the higher degree cases open. However, we can use the main theorem to exhibit (for every $n \geq 2$) a polynomial of degree n whose Galois group is S_n and so is not solvable if $n \geq 5$. We devote the rest of this section to this piece of classical algebra.

A key aspect of Galois theory is that if $E \supseteq F$ is a splitting field of the polynomial f in $F[x]$, then each automorphism in $\text{gal}(E : F)$ permutes the roots of f . Moreover, the coefficients of f are functions of the roots that remain unchanged when the roots are permuted. For example, if the roots are u_1, u_2 , and u_3 ,

$$\begin{aligned} f &= (x - u_1)(x - u_2)(x - u_3) \\ &= x^3 - (u_1 + u_2 + u_3)x^2 + (u_1u_2 + u_1u_3 + u_2u_3)x - u_1u_2u_3. \end{aligned} \quad (1)$$

We formalize this idea as follows.

If F is a field and $F[x_i] = F[x_1, x_2, \dots, x_n]$ is the polynomial ring in n indeterminates x_1, x_2, \dots, x_n , a polynomial $s(x_i) = s(x_1, x_2, \dots, x_n)$ in $F[x_i]$ is called **symmetric** if

$$s(x_{\sigma 1}, x_{\sigma 2}, \dots, x_{\sigma n}) = s(x_1, x_2, \dots, x_n) \quad \text{for all } \sigma \in S_n.$$

Thus, $x_1^2x_2x_3 + x_1x_2^2x_3 + x_1x_2x_3^2$ and $x_1^3 + x_2^3 + x_3^3$ are symmetric polynomials in $F[x_1, x_2, x_3]$. The coefficients of the polynomial (1) are symmetric polynomials in the u_i , and these polynomials play an important role in what we do next.

Given $F[x_i] = F[x_1, x_2, \dots, x_n]$, the **elementary symmetric polynomials** $s_0, s_1, s_2, \dots, s_n$ in $F[x_i]$ are defined as follows:

$$\begin{aligned} s_0 &= s_0(x_i) = 1, \\ s_k &= s_k(x_i) = \sum_{i_1 < i_2 < \dots < i_k} x_{i_1}x_{i_2} \cdots x_{i_k} \quad \text{for } k = 1, 2, \dots, n. \end{aligned}$$

Thus, $s_1 = x_1 + x_2 + \cdots + x_n$ and $s_n = x_1 x_2 \cdots x_n$ for any n . If $n = 3$, then

$$\begin{aligned}s_0 &= s_0(x_1, x_2, x_3) = 1, \\s_1 &= s_1(x_1, x_2, x_3) = x_1 + x_2 + x_3, \\s_2 &= s_2(x_1, x_2, x_3) = x_1 x_2 + x_1 x_3 + x_2 x_3, \\s_3 &= s_3(x_1, x_2, x_3) = x_1 x_2 x_3.\end{aligned}$$

Hence, (1) is the case for $n = 3$ of the easily verified formula

$$(x - u_1)(x - u_2) \cdots (x - u_n) = x^n - s_1(u_i)x^{n-1} + s_2(u_i)x^{n-2} + \cdots + (-1)^ns_n(u_i). \quad (2)$$

We discussed this material at length in Section 4.5, and our interest here is in the use of (2) to calculate a certain Galois group.

If x_1, \dots, x_n are indeterminates over a field F , let $E = F(x_1, \dots, x_n)$ denote the field of **rational forms** over F ; that is, E is the field of quotients of the integral domain $F[x_i] = F[x_1, \dots, x_n]$. Hence, the elements of E are quotients

$$r(x_i) = \frac{f(x_i)}{g(x_i)}.$$

where f and g are polynomials in the variables x_1, x_2, \dots, x_n , and $g \neq 0$. If $\sigma \in S_n$, we define

$$\bar{\sigma} : E \rightarrow E \quad \text{by} \quad \bar{\sigma}[r(x_1, \dots, x_n)] = r(x_{\sigma 1}, \dots, x_{\sigma n}).$$

One verifies that $\bar{\sigma}$ is an automorphism of E that fixes F ; that is, $\bar{\sigma} \in \text{gal}(E : F)$. Moreover, $\sigma \mapsto \bar{\sigma}$ is a one-to-one group homomorphism $S_n \rightarrow \text{gal}(E : F)$. Write its image as $\bar{S}_n = \{\bar{\sigma} \mid \sigma \in S_n\}$ so that \bar{S}_n is a subgroup of $\text{gal}(E : F)$ isomorphic to S_n . Our interest is in the fixed field S of \bar{S}_n in E :

$$\begin{aligned}S &= E_{\bar{S}_n} \\&= \{r(x_i) \mid \bar{\sigma}(r) = r \text{ for all } \sigma \in S_n\} \\&= \{r(x_1, \dots, x_n) \mid r(x_{\sigma 1}, \dots, x_{\sigma n}) = r(x_1, \dots, x_n) \text{ for all } \sigma \in S_n\}.\end{aligned}$$

This is called the field of **symmetric rational forms** over F . Note that the Dedekind–Artin theorem gives

$$[E : S] = |\bar{S}_n| = n!. \quad (3)$$

In fact, we claim that $E \supseteq S$ is Galois and $S = F(s_0, s_1, \dots, s_n) = F(s_j)$, where s_0, s_1, \dots, s_n are the elementary symmetric polynomials in the variables x_i . Clearly,

$$F(s_j) \subseteq S \subseteq E.$$

If t is an indeterminate over E , consider the polynomial $f(t)$ in $F(s_j)[t]$ given by

$$f(t) = (t - x_1)(t - x_2) \cdots (t - x_n) = t^n - s_1 t^{n-1} + \cdots + (-1)^n s_n.$$

Then E is the splitting field of $f(t)$ over $F(s_j)$ and so, as $\deg f = n$, we have $[E : F(s_j)] \leq n!$ by Theorem 2 §6.3. This result, along with (3), shows that $S = F(s_j)$. Moreover, $f(t)$ is clearly separable over $S = F(s_j)$, so $E \supseteq S$ is a Galois extension (by Theorem 3 §10.2) and $|\text{gal}(E : S)| = [E : S] = n!$. Because $\bar{S}_n \subseteq \text{gal}(E : S)$ has also order $n!$, this gives $\text{gal}(E : S) = \bar{S}_n$. The next theorem collects these results.

Theorem 3. *Let F be a field, let $E = F(x_1, \dots, x_n)$ be the field of rational forms over F , and let $S \subseteq E$ be the subfield of all symmetric rational forms.*

- (1) $E \supseteq S$ is a Galois extension, $[E : S] = n!$, and $\text{gal}(E : S) \cong S_n$.
- (2) $S = F(s_0, s_1, s_2, \dots, s_n)$, where s_j are the elementary symmetric polynomials in the x_i , and E is the splitting field over S of the polynomial $f(t) = t^n - s_1 t^{n-1} + s_2 t^{n-2} + \dots + (-i)^n s_n$.

Corollary 1. If $n \geq 5$, a nonsolvable polynomial of degree n exists.

Proof. The polynomial $f(t) \in S[t]$ in Theorem 3 is not solvable over S because its Galois group S_n is not solvable if $n \geq 5$. ■

The fact that $S = F(s_0, s_1, \dots, s_n)$ in Theorem 3 means that every symmetric rational form in the variables x_1, x_2, \dots, x_n is the quotient of two polynomials in the elementary symmetric polynomials s_0, s_1, \dots, s_n in these variables. In fact, every symmetric polynomial in the x_i is actually a polynomial in the s_j , a fact proved (without any field theory) in Theorem 4 §4.5.

Because every group of order n is isomorphic to a subgroup of S_n , part (4) of the main theorem provides a bonus.

Corollary 2. Every finite group is isomorphic to the Galois group of a finite Galois extension.

Surprisingly, no one knows whether every finite group is isomorphic to the Galois group of a finite Galois extension of \mathbb{Q} . Even small order groups can be complicated. For example,¹²⁶ the quaternion group Q is the Galois group of $E \supseteq \mathbb{Q}$, where E is the splitting field of $x^8 - 72x^6 + 180x^4 - 144x^2 + 36$. On the other hand, in 1956 the Russian mathematician I. R. Shafarevich proved that if $F \supseteq \mathbb{Q}$ is a finite algebraic extension, and G is any finite solvable group, then $G \cong \text{gal}(E : F)$ for some Galois extension $E \supseteq F$.

Exercises 10.3

1. Find a radical extension of \mathbb{Q} containing

(a) $\sqrt{3}(\sqrt[3]{5} - \sqrt[5]{7})$	(b) $(\sqrt{5} - 3)(4 - 3\sqrt[5]{6})$
---	--
2. In each case, show that f is not solvable by radicals.

(a) $f = x^5 - 4x - 2$	(b) $f = x^5 - 6x^2 + 2$
------------------------	--------------------------
3. Show that $x^7 - 14x + 2$ in $\mathbb{Q}[x]$ has Galois group S_7 .
4. If p is a prime and $f \in \mathbb{Q}[x]$ is irreducible of degree p , and if f has exactly two nonreal roots, show that f has Galois group S_p .
5. Show that every polynomial of degree at most 4 is solvable. [Hint: Theorem 3 §9.2.]
6. If f is a separable, irreducible cubic in $F[x]$, F a field, show that its Galois group is S_3 or C_3 . [Hint: Theorem 3 §10.1.]
7. Consider $f = x^3 - 3x + 1$ in $\mathbb{Q}[x]$. Find the roots of f and determine the Galois group.
8. Let $f \in F[x]$, where F is a field and $\text{char } F \neq 2$. Assume that $\deg f = n$ and that the roots u_1, u_2, \dots, u_n in a splitting field $E \supseteq F$ are distinct. If $G = \text{gal}(E : F)$, view G as a group of permutations in S_X , where $X = \{u_1, \dots, u_n\}$. [See Theorem 3 §10.1.]

¹²⁶Dean, R. A., *American Mathematical Monthly*, 88 (1981), 42–45.

Define $\Delta \in E$ by $\Delta = \prod_{i < j} (u_i - u_j)$, so if $n = 3$, $\Delta = (u_1 - u_2)(u_1 - u_3)(u_2 - u_3)$. The element Δ^2 is called the **discriminant** of f .

- (a) Show that $\Delta^2 \in F$.
- (b) Show that the permutation $\sigma \in S_X$ is even if and only if $\sigma(\Delta) = \Delta$ and that σ is odd if and only if $\sigma(\Delta) = -\Delta$.
- (c) Show that $F(\Delta)$ corresponds to the even permutations in S_X in the Galois correspondence.
- (d) Show that G consists of even permutations if and only if $\Delta \in F$.
- (e) If $f = x^2 + bx + c$, show that $\Delta^2 = b^2 - 4c$, the usual discriminant.
- (f) If $f = x^3 + bx + c$, show that $\Delta^2 = -4b^3 - 27c^2$ is the usual discriminant. [Hint: $u_1 + u_2 + u_3 = 0$ and $u_1u_2 + u_1u_3 + u_2u_3 = b$ imply that $(u_i - u_j)^2 = -b - 3u_iu_j$.]

10.4 CYCLOTOMIC POLYNOMIALS AND WEDDERBURN'S THEOREM

If n is a positive integer, the irreducible factors of $x^n - 1$ in $\mathbb{Q}[x]$ are called cyclotomic polynomials and are important in number theory. In this section, we derive several properties of these polynomials using Galois theory, and use them to prove a famous theorem of Wedderburn: Every finite division ring is a field.

If F is a field and $n \geq 1$ is an integer, we let $E \supseteq F$ be a splitting field of $x^n - 1$. The roots form a subgroup of E^* , which is cyclic by Theorem 7 §6.4, and this group has order n if and only if $\text{char } F$ does not divide n (Theorem 1 §10.3).¹²⁷ In this case, a generator of this group is called a primitive n th root of unity over F . This group has exactly $\varphi(n)$ generators, where φ is the Euler function (Section 2.6).

Let $w_1, w_2, \dots, w_{\varphi(n)}$ be the primitive n th roots of unity over a field F (where $\text{char } F$ does not divide n) and define

$$\Phi_n = (x - w_1)(x - w_2) \cdots (x - w_{\varphi(n)}).$$

This is called the n th **cyclotomic polynomial** over F . Clearly, $\Phi_n \in E[x]$; in fact, $\Phi_n \in F[x]$ —see Theorem 1.

Example 1. $\Phi_1 = x - 1$ over any field.

Example 2. If $\text{char } F \neq 2$, show that $\Phi_4 = x^2 + 1$.

Solution. Because $x^4 - 1 = (x - 1)(x + 1)(x^2 + 1)$, the splitting field of $x^4 - 1$ is $F(w)$, where $w^2 = -1$. Hence, the primitive fourth roots of unity are w and $-w$, so $\Phi_4 = (x - w)(x + w) = x^2 + 1$. \square

Example 3. Show that $\Phi_p = x^{p-1} + x^{p-2} + \cdots + x + 1$, where $p \neq \text{char } F$ is a prime.

Solution. The p th roots of unity are the (distinct) roots of $x^p - 1$ and every one (except 1) is primitive because p is a prime. As $x^p - 1 = (x - 1)\Phi_p$, the result follows. \square

Note: For the rest of this section, we adopt the convention that if $n \geq 1$ is an integer, $d|n$ means that d is a *positive* divisor of n .

¹²⁷This is intended to include the case $\text{char } F = 0$.

Theorem 1. Let F be a field, where $\text{char } F$ does not divide n .

- (1) $\Phi_n \in F[x]$.
- (2) $x^n - 1 = \prod_{d|n} \Phi_d$.

Proof. Let w be any primitive n th root of unity over F .

(1) If $E = F(w)$, then $\Phi_n \in E[x]$ is clear. If $\sigma \in \text{gal}(E : F)$, then σ permutes the primitive n th roots of unity and so fixes every coefficient of Φ_n . But $E \supseteq F$ is Galois (by Theorem 1 §10.3), so these coefficients are in F .

(2) If $d|n$, the primitive d th roots of unity are precisely the elements of order d in $U = \langle w \rangle$. On the other hand, every element of U is a primitive d th root of unity for a unique positive divisor d of n . Thus,

$$x^n - 1 = \prod_{u \in U} (x - u) = \prod_{d|n} [\prod_{u \in U, o(u)=d} (x - u)] = \prod_{d|n} \Phi_d. \quad \blacksquare$$

If p is a prime, (2) gives $x^p - 1 = \Phi_p \Phi_1 = (x - 1) \Phi_p$. This in turn gives

$$\Phi_p = x^{p-1} + x^{p-2} + \cdots + x + 1, \quad p \text{ any prime}$$

as in Example 3. In general, (2) in Theorem 1 gives a recursive method for determining the polynomials Φ_n :

$$\begin{aligned} \Phi_4 &= \frac{x^4 - 1}{\Phi_1 \Phi_2} = \frac{x^4 - 1}{(x-1)(x+1)} = x^2 + 1 \\ \Phi_6 &= \frac{x^6 - 1}{\Phi_1 \Phi_2 \Phi_3} = \frac{x^6 - 1}{(x-1)(x+1)(x^2+x+1)} = x^2 - x + 1. \end{aligned}$$

Note that all the coefficients of the Φ_n are integers. Over \mathbb{Q} this holds in general.

Theorem 2. The cyclotomic polynomials Φ_n over \mathbb{Q} have integral coefficients.

Proof. Use induction on n , beginning with $\Phi_1 = x - 1$. In general, (2) of Theorem 1 gives $x^n - 1 = \Phi f$, where

$$f = \prod_{d|n, d \neq n} \Phi_d$$

has integer coefficients by induction. Also, f is monic (because each Φ_d is monic), so $x^n - 1 = fq + r$ in $\mathbb{Z}[x]$, where either $r = 0$ or $\deg r < \deg f$. But then $r = (\Phi_n - q)f$ forces $r = 0$ and so $\Phi_n = q \in \mathbb{Z}[x]$. \blacksquare

It is interesting to note that $n = 105$ is the smallest value of n for which Φ_n has an integer coefficient other than 0, 1, or -1 .

We can now prove a famous theorem of J. H. M. Wedderburn. He proved it in 1905,¹²⁸ but the proof we give is due to Ernst Witt in 1931. It utilizes the class equation for a finite group given in Section 8.2 and also requires two preliminary results, the first of which is an easy consequence of the definition of the cyclotomic polynomials.

Lemma 1. If $d|n$, then Φ_n divides $(x^n - 1)/(x^d - 1)$ in $\mathbb{Z}[x]$.

Proof. Observe that Φ_n divides

$$x^n - 1 = (x^d - 1) \times \left(\frac{x^n - 1}{x^d - 1} \right),$$

¹²⁸Wedderburn, J.H.M., A Theorem on finite algebras, Transactions of the American Mathematical Society, 6 (1905), 349–352.

so it suffices to show that Φ_n and $x^d - 1$ are relatively prime. But this follows because the roots in \mathbb{C} of Φ_n are primitive n th roots of unity and so cannot be roots of $x^d - 1$. \blacksquare

Lemma 2. Suppose $q^d - 1$ divides $q^n - 1$, where $q > 1$, $d \geq 1$, and $n \geq 1$. Then $d|n$.

Proof. Write $n = ad + r$ in \mathbb{Z} , where $0 \leq r < d$. Then, working modulo $q^d - 1$, $1 \equiv q^n \equiv q^{ad} \cdot q^r \equiv q^r$, which implies that $r = 0$. \blacksquare

Theorem 3. Wedderburn's Theorem. Every finite division ring is a field.

Proof. If R is a finite division ring, let $Z = \{z \in R \mid zr = rz \text{ for all } r \in R\}$ denote its center. Then Z is a finite field, say $|Z| = q$. If $\{r_1, \dots, r_n\}$ is a basis of R as a vector space over Z , then $|R| = q^n$. We consider the group R^* and its center $Z(R^*) = Z^*$. Clearly, $|R^*| = q^n - 1$ and $|Z^*| = q - 1$. If R is not commutative, then $n > 1$, so the class equation for R^* (Theorem 3 §8.2) reads as follows: If class u_1 , class u_2, \dots , class u_m are the nonsingleton conjugacy classes in R^* , then

$$|R^*| = |Z^*| + \sum_{i=1}^m |R^* : N(u_i)|. \quad (*)$$

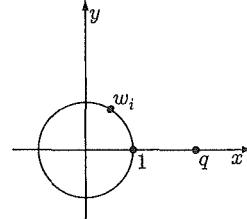
Now $R_i = \{r \in R \mid ru_i = u_i r\}$ is a division subring of R that contains Z and so has order $|R_i| = q^{d_i}$ for some d_i . Moreover, $N(u_i) = R_i^*$, so $|N(u_i)| = q^{d_i} - 1$, which divides $|R^*| = q^n - 1$ by Lagrange's theorem. Hence, $d_i|n$ by Lemma 2, so

$$|R^* : N(u_i)| = \frac{q^n - 1}{q^{d_i} - 1}$$

is a multiple of $\Phi_n(q)$ by Lemma 1 because $|R^* : N(u_i)| = |\text{class } u_i| > 1$. Because $\Phi_n(q)$ also divides $|R^*| = q^n - 1$, $(*)$ implies that $\Phi_n(q)$ divides $|Z^*| = q - 1$. But if w_1, w_2, \dots are the primitive n th roots of unity in \mathbb{C} , then $w_i \neq 1$ for each i because $n > 1$. Hence, $|q - w_i| > (q - 1)$ for each i (see the diagram), so

$$|\Phi_n(q)| = \prod_i |q - w_i| > (q - 1)^t \geq q - 1.$$

This contradiction establishes Wedderburn's theorem. \blacksquare



Wedderburn's theorem can be extended. If R is a finite division ring, say $|R| = n$, Lagrange's theorem shows that $r^{n-1} = 1$ for all $r \neq 0$ in R , so $r^n = r$ for all $r \in R$. Hence, finite division rings are periodic, where R is called a **periodic ring** if, for all $r \in R$, $r^n = r$ for some integer n (depending on r). Thus, Wedderburn's theorem is a special case of Jacobson's theorem: *Every periodic ring is commutative*.

Returning to cyclotomic polynomials, we conclude this section by showing that Φ_n is irreducible in $\mathbb{Q}[x]$ for every $n \geq 1$ (so the factorization of $x^n - 1$ into irreducibles in $\mathbb{Q}[x]$ is given by (2) of Theorem 1). This is true if n is prime by Example 13 §4.2, but to prove it in general, we need some notions from Chapter 5.

If $f \neq 0$ is a polynomial in $\mathbb{Z}[x]$, the gcd of its coefficients is called the *content* of f , denoted $c(f)$, and f is said to be *primitive* if $c(f) = 1$. We can easily show (Lemma 4 §5.1) that if $c(f) = c$, then $f = cf_1$, where f_1 is primitive. The key observation

about this is Gauss' lemma (Theorem 8 §5.1), which states that $c(fg) = c(f)c(g)$ for all nonzero f and g in $\mathbb{Z}[x]$.

Lemma 3. Let $f \in \mathbb{Z}[x]$ be monic. If $f = p_1 q_1$ in $\mathbb{Q}[x]$, then f can be written as $f = pq$ in $\mathbb{Z}[x]$, where p and q are monic and $p = rp_1$ and $q = sq_1$ for some r and s in \mathbb{Q} .

Proof. Choose $a, b \in \mathbb{Z}$ such that $ap_1 = p_0$ and $bq_1 = q_0$ are in $\mathbb{Z}[x]$ and then write $p_0 = cp$ and $q_0 = dq$, where $p, q \in \mathbb{Z}[x]$ are primitive. Hence, $abf = cdpq$. Because f is also primitive (being monic), Gauss' lemma gives $ab = c(abf) = cd$. Hence, $f = pq$ in $\mathbb{Z}[x]$, so, as f is monic, p and q may be assumed to be monic. As $ap_1 = cp$ and $bq_1 = dq$, the proof is complete. \blacksquare

Theorem 4. Φ_n is irreducible in $\mathbb{Q}[x]$ for every n .

Proof. Let w be a primitive n th root of unity. Because Φ_n is monic in $\mathbb{Z}[x]$ and $\Phi_n(w) = 0$, the minimal polynomial m_1 of w over \mathbb{Q} divides Φ_n . Hence, by Lemma 3, write $\Phi_n = mf$ in $\mathbb{Z}[x]$, where m is monic, $m(w) = 0$, and m is irreducible in $\mathbb{Q}[x]$. We demonstrate that $\deg m = \deg \Phi_n$ by showing that $m(w^k) = 0$ for all integers k relatively prime to n (every primitive root of unity is such a w^k).

To do so, it suffices to show that $m(w^p) = 0$ for any prime p not dividing n . Suppose that $m(w^p) \neq 0$ for such a prime. Then $0 = \Phi_n(w^p) = m(w^p)f(w^p)$, so $f(w^p) = 0$. Thus, w is a root of $f(x^p)$, so (as m is irreducible) $f(x^p) = mg$ for $g \in \mathbb{Q}[x]$. But m is monic in $\mathbb{Z}[x]$, so $f(x^p) = qm + r$ in $\mathbb{Z}[x]$ where either $r = 0$ or $\deg r < \deg m$. This gives $r = (g - q)m$, so $g = q \in \mathbb{Z}[x]$. Hence, $f(x^p) = mg$ holds in $\mathbb{Z}[x]$, and so taking the coefficients modulo p ,

$$\bar{m} \bar{g} = \bar{f}(x^p) = \bar{f}(x)^p \quad \text{in } \mathbb{Z}_p[x].$$

Thus, \bar{m} and \bar{f} have a common irreducible factor in $\mathbb{Z}_p[x]$. On the other hand, $x^n - 1 = \Phi_n h$ for some $h \in \mathbb{Z}[x]$, so

$$x^n - 1 = \bar{m} \bar{f} \bar{h} \quad \text{in } \mathbb{Z}_p[x].$$

Hence, $x^n - 1$ has a multiple zero in $\mathbb{Z}_p[x]$, a contradiction as p does not divide n . \blacksquare

Example 4. Note that it is essential that Φ_n is taken over \mathbb{Q} in Theorem 4. For example, $\Phi_6 = x^2 - x + 1$ becomes $\Phi_6 = (x + 1)^2$ in $\mathbb{Z}_3[x]$.

Exercises 10.4

1. Find (a) Φ_8 , (b) Φ_{10} , (c) Φ_{12} , (d) Φ_{15} , and (e) Φ_{18} .
2. If p is a prime, show that $\Phi_{p^n} = 1 + x^q + x^{2q} + \cdots + x^{(p-1)q}$, where $q = p^{n-1}$.
3. If $n \geq 3$ is odd, show that $\Phi_{2n} = \Phi_n(-x)$. [Hint: $\Phi_2(x) = -\Phi_1(-x)$.]
4. Show that $n = \sum_{d|n} \varphi(d)$ if $n \geq 1$ (φ is the Euler function). [Hint: Theorem 1(2).]
5. If $\gcd(m, n) = 1$, prove that $x^{mn} - 1$ and $(x^m - 1)(x^n - 1)$ have the same splitting fields over \mathbb{Q} .
6. (a) Show finite subrings of division rings are division rings.
 (b) If R is a division ring of characteristic $p \neq 0$, show that any finite subgroup G of R^* is cyclic. Is this true if $\text{char } R = 0$? [Hint: Regard $\mathbb{Z}_p \subseteq R$ and if $G = \{g_1, g_2, \dots, g_n\}$, consider $R_0 = \left\{ \sum_{i=1}^n r_i g_i \mid r_i \in \mathbb{Z}_p \right\}$.]

7. The Möbius function $\mu : \mathbb{Z}^+ \rightarrow \{0, 1, -1\}$ is defined by

$$\mu(1) = 1,$$

$\mu(n) = 0$ if $n = p^2m$ for some prime p ,

$\mu(n) = (-1)^k$ if $n = p_1p_2 \cdots p_k$, where p_1, \dots, p_k are distinct primes.

Show that $\sum_{d|n} \mu(d) = \begin{cases} 1, & \text{if } n = 1, \\ 0, & \text{if } n > 1. \end{cases}$

8. (a) Let maps $\alpha : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ and $\beta : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$ be related by $\alpha(n) = \sum_{d|n} \beta(d)$. Show that β is given in terms of α by

$$\beta(n) = \sum_{d|n} \mu\left(\frac{n}{d}\right) \alpha(d) = \sum_{d|n} \mu(d) \alpha\left(\frac{n}{d}\right).$$

This is called the **Möbius inversion formula**. [Hint: $d|n$ and $c|(n/d) \Leftrightarrow dc|n \Leftrightarrow c|n$ and $d|(n/c)$.]

(b) Prove that $\Phi_n = \prod_{d|n} (x^d - 1)^{\mu(n/d)}$, where μ is the Möbius μ -function.

[Hint: Exercise 7; use a formal logarithm.]

9. Let $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, where p_i are distinct primes, and let $m = p_1 p_2 \cdots p_r$. Show that $\Phi_n(x) = \Phi_m(x^{n/m})$. [Hint: Exercise 8(b).]

10. If p is an odd prime and p does not divide n , show that $\Phi_{np}(x) \cdot \Phi_n(x) = \Phi_n(x^p)$.

[Hint: Exercise 8(b).]

Chapter 11

Finiteness Conditions for Rings and Modules

A scientist worthy of the name, above all a mathematician, experiences in his work the same impression as an artist; his pleasure is as great and of the same nature.

—Henri Poincaré

The field \mathbb{C} of complex numbers is a two-dimensional vector space over \mathbb{R} . A ring that is also a vector space over a field F is called an algebra over F . In the nineteenth century many attempts were made to describe the division algebras that are finite dimensional over \mathbb{R} , and in particular those (like \mathbb{C}) that are fields. Certainly the three-dimensional examples would have had applications to physics, but after looking in vain for such an algebra, the first success came in 1843 when W.R. Hamilton discovered the ring \mathbb{H} of quaternions, a four-dimensional algebra that, surprisingly, was not commutative. It was not until 1878 that G. Frobenius showed that there is *no* three-dimensional example, and that the only possible associative examples are \mathbb{R} , \mathbb{C} , and \mathbb{H} . Meanwhile, G. Grassmann described many examples that were rings but not necessarily division rings, of which the matrix algebras, constructed in 1858 by A. Cayley, were an important special case. The next event in the development of the theory came in 1907 when J.H.M. Wedderburn gave the first characterization of the simple finite dimensional algebras. Finally, in 1927, E. Artin extended Wedderburn's theorem to a result about rings by eliminating the dependence on the fact that the ring is finite dimensional as an algebra, replacing it with finiteness conditions on the set of left ideals. These seminal results mark the beginning of the theory of noncommutative ring theory, and we prove them in this chapter.

11.1 WEDDERBURN'S THEOREM

If F is a field, a ring R is called an *F -algebra* if it is a vector space over F and $t(ab) = a(tb)$ for all $t \in F$ and all $a, b \in R$. Hence, \mathbb{C} is a two-dimensional \mathbb{R} -algebra and the ring $M_n(F)$ of all $n \times n$ matrices over F is an n^2 -dimensional F -algebra. The ring \mathbb{H} of real quaternions is a four-dimensional algebra over \mathbb{R} that has the distinction of being a division ring (every nonzero element is a unit). Wedderburn's theorem asserts that if R is a simple, finite dimensional algebra then $R \cong M_n(D)$ for some $n \geq 1$ and some division ring D . We are going to prove an extension of this theorem that removes the restriction that R is an algebra. The first task is to find an appropriate "finiteness condition" on R to replace the finite dimensional requirement.

Finiteness Conditions

If R is a ring, recall (Section 7.1) that a left R -module is an additive abelian group with a left R -action $rx \in M$, $r \in R$, $x \in M$, such that the axioms for a vector space are satisfied. In the nineteenth and early twentieth centuries, most of the modules studied were finite dimensional vector spaces. Emmy Noether, and later Emil Artin, realized that the right way to extend this finite dimensional condition was to use finiteness conditions on the set of submodules. Lemma 1 below identifies the most important of these.

If S is a nonempty set of submodules of a module M , then $N \in S$ is said to be **maximal** in S if $N \subseteq K \in S$ implies that $K = N$. Similarly N is called **minimal** in S if $K \subseteq N \in S$ implies that $K = N$. For example, the maximal ideals in a ring are the maximal members of $S = \{A \mid A \text{ is an ideal of } R \text{ and } A \neq R\}$.

Lemma 1. *If M is a module, consider the following conditions where the K_i denote submodules of M . Then $(\text{ACC}) \Leftrightarrow (\text{MAX})$ and $(\text{DCC}) \Leftrightarrow (\text{MIN})$.*

- (ACC) *If $K_1 \subseteq K_2 \subseteq \dots$ then $K_n = K_{n+1} = \dots$ for some n .*
- (MAX) *Every nonempty set of submodules of M has a maximal member.*
- (DCC) *If $K_1 \supseteq K_2 \supseteq \dots$ then $K_n = K_{n+1} = \dots$ for some n .*
- (MIN) *Every nonempty set of submodules of M has a minimal member.*

Proof. If (MAX) holds then (ACC) holds where K_n is any maximal member of $\{K_i \mid i \geq 1\}$. Conversely, suppose that S is a nonempty set of submodules of M that has no maximal member. Choose $K_1 \in S$. Since K_1 is not maximal in S there exists $K_2 \in S$ such that $K_1 \subset K_2$. But K_2 is also not maximal in S so we obtain $K_1 \subset K_2 \subset K_3$ for some $K_3 \in S$. This process continues indefinitely¹²⁹ to violate (ACC). This proves that $(\text{ACC}) \Leftrightarrow (\text{MAX})$; the proof that $(\text{DCC}) \Leftrightarrow (\text{MIN})$ is analogous. ■

The notations ACC and DCC refer to the **ascending** and **descending chain conditions**, on submodules, respectively. A module M is called **noetherian** if it has the ACC, and M is called **artinian** if it has the DCC. If R is an F -algebra with unity 1 then any left module $_RM$ becomes an F -space via the action $tx = (t1)x$

¹²⁹Actually this requires a set-theoretical theorem called transfinite recursion. This is discussed in Appendix D.

for all $t \in F$ and $x \in M$. Hence every submodule of M is a subspace, so finite dimensional modules are both noetherian and artinian.¹³⁰

Example 1. Regarded as a module over itself, \mathbb{Z} is noetherian but not artinian.

Proof. \mathbb{Z} is not artinian as $\mathbb{Z} \supset \mathbb{Z}2 \supset \mathbb{Z}4 \supset \dots$ shows. For the converse, suppose $K_1 \subseteq K_2 \subseteq \dots$ are subgroups of \mathbb{Z} . Then $K = \bigcup K_i$ is a subgroup and so $K = \mathbb{Z}k$ for some $k \in \mathbb{Z}$ by Theorem 7 §2.4. But then $k \in K_n$ for some n and it follows easily that $K_n = K_{n+1} = \dots = K$. Hence \mathbb{Z} is noetherian. \square

If $p \in \mathbb{Z}$ is a prime, let $X = \{\frac{m}{p^k} \in \mathbb{Q} \mid m \in \mathbb{Z}, k \geq 0\}$. This is an additive subgroup of \mathbb{Q} , and one verifies that the groups $\mathbb{Z} \subset \mathbb{Z}\frac{1}{p} \subset \mathbb{Z}\frac{1}{p^2} \subset \dots \subset X$ are the *only* subgroups of X containing \mathbb{Z} (Exercise 9). The factor group $\mathbb{Z}_{p^\infty} = X/\mathbb{Z}$ is called the **Prüfer p -group**, and this shows that the only subgroups of \mathbb{Z}_{p^∞} are

$$0 \subset \mathbb{Z}x_1 \subset \mathbb{Z}x_2 \subset \mathbb{Z}x_3 \subset \dots \subset \mathbb{Z}_{p^\infty},$$

where $x_k = \frac{1}{p^k} + \mathbb{Z}$ for each k . Furthermore, $o(x_k) = p^k$ and $px_{k+1} = x_k$ hold for each k . Hence \mathbb{Z}_{p^∞} is an infinite group, but every proper subgroup is finite. Clearly

Example 2. The Prüfer group \mathbb{Z}_{p^∞} is artinian but not noetherian as a \mathbb{Z} -module.

The following basic properties of the ACC and the DCC will be needed.

Lemma 2. *If $N \subseteq M$ are modules then M is noetherian (artinian) if and only if the same is true of both N and M/N .*

Proof. We prove the noetherian case; the other is analogous. If M has the ACC then N has the ACC because submodules of N are submodules of M . On the other hand, every submodule X of M/N has the form $X = K/N$ for some submodule K of M containing N by Theorem 5 §8.1. Hence every ascending chain in M/N takes the form $K_1/N \subseteq K_2/N \subseteq \dots$ where $K_1 \subseteq K_2 \subseteq \dots$ in M . It follows that M/N has the ACC.

Conversely, let $K_1 \subseteq K_2 \subseteq \dots$ be a chain from M . If both N and M/N have the ACC then the chains $N \cap K_1 \subseteq N \cap K_2 \subseteq \dots$ and $\frac{N+K_1}{N} \subseteq \frac{N+K_2}{N} \subseteq \dots$ both terminate, so there exists $n \geq 1$ such that $N \cap K_n = N \cap K_{n+1} = \dots$ and $\frac{N+K_n}{N} = \frac{N+K_{n+1}}{N} = \dots$ (whence $N + K_n = N + K_{n+1} = \dots$). Since $K_i \subseteq K_{i+1}$ for each i , we have $K_{i+1} \cap (K_i + N) = K_i + (K_{i+1} \cap N)$ by the modular law (Theorem 2 §7.1). So if $i \geq n$ we have

$$\begin{aligned} K_{i+1} &= K_{i+1} \cap (K_{i+1} + N) = K_{i+1} \cap (K_i + N) = K_i + (K_{i+1} \cap N) \\ &= K_i + (K_i \cap N) = K_i. \end{aligned}$$

This proves that M is noetherian. \blacksquare

Corollary. *A sum $M = M_1 + \dots + M_n$ of modules is artinian (noetherian) if and only if the same is true of each M_i .*

Proof. We prove only the artinian case, the other being analogous. If M is artinian, so are its submodules M_i by Lemma 2. Conversely, if $n \geq 2$, assume inductively that $K = M_2 + \dots + M_n$ is artinian. By Corollary 2 of Theorem 1 §7.1, it follows

¹³⁰The ACCP designation used extensively in Chapter 5 is just the ACC applied to the set of all principal ideals in an integral domain.

that $M/K = (M_1 + K)/K \cong M_1/(M_1 \cap K)$ is also artinian by Lemma 2, being an image of M_1 . Hence, M is artinian, again by Lemma 2, as required. \blacksquare

Let M_1, \dots, M_n be modules. The external direct sum $M = M_1 \oplus \dots \oplus M_n$ is artinian (noetherian) if and only if the same is true of each M_i by the Corollary because M is a sum of submodules M'_i isomorphic to the M_i .

Because the \mathbb{Z} -action on an abelian group is naturally written on the left, we discussed only left R -modules in Chapter 7. However, an additive abelian group M is called a **right R -module** (written M_R) if R acts on the right: That is, for any $r \in R$ and $x \in M$, an element $xr \in M$ is defined such that

$$(x+y)r = xr + yr, \quad x(r+s) = xr + xs, \quad (xr)s = x(rs), \quad \text{and} \quad x1 = x$$

hold for all $x, y \in M$ and all $r, s \in R$.¹³¹ With this, the definition of submodules and homomorphisms of right modules are analogous to those for left modules. Moreover, the analogues of theorems about left modules in Section 7.1 go through verbatim for right modules. Note that the distinction does not matter for a commutative ring R because a left R -module M becomes a right module if we define the action by $x \cdot r = rx$ for all $r \in R$ and $x \in M$.

A ring R is called **left artinian (left noetherian)** if ${}_R R$ is artinian (noetherian), with similar definitions on the right. The submodules of ${}_R R$ (R_R) are the left (right) ideals.

Example 3. In each case consider the subring R of $M_2(\mathbb{R})$.

(1) $R = \begin{bmatrix} \mathbb{Z} & 0 \\ \mathbb{Q} & \mathbb{Q} \end{bmatrix}$ is left noetherian but not right noetherian.

(2) $R = \begin{bmatrix} \mathbb{Q} & 0 \\ \mathbb{R} & \mathbb{R} \end{bmatrix}$ is left artinian but not right artinian.

Solution.

(1) If $A = \begin{bmatrix} 0 & 0 \\ \mathbb{Q} & 0 \end{bmatrix}$ then A is an ideal of R and $R/A \cong \mathbb{Z} \times \mathbb{Q}$ is noetherian as a ring by the Corollary to Lemma 2 (verify). Hence, R/A is noetherian as a left R -module (the left R -submodules of R/A are just the left ideals of the ring R/A). Since ${}_R A$ is also noetherian (verify), it follows by Lemma 2 that R is left noetherian. But the sequence $\begin{bmatrix} 0 & 0 \\ \mathbb{Z} & 0 \end{bmatrix} \subset \begin{bmatrix} 0 & 0 \\ \frac{1}{2}\mathbb{Z} & 0 \end{bmatrix} \subset \begin{bmatrix} 0 & 0 \\ \frac{1}{4}\mathbb{Z} & 0 \end{bmatrix} \subset \dots$ of right ideals shows that R is not right noetherian. This proves (1).

(2) Observe that \mathbb{R} is not finite dimensional as a \mathbb{Q} -space (otherwise \mathbb{R} would be countable contradicting Cantor's theorem from set theory). Hence, there exist \mathbb{Q} -subspaces $X_1 \supset X_2 \supset \dots$ in \mathbb{R} . But then $\begin{bmatrix} 0 & 0 \\ X_1 & 0 \end{bmatrix} \supset \begin{bmatrix} 0 & 0 \\ X_2 & 0 \end{bmatrix} \supset \dots$ are right ideals of R , proving that R is not right artinian. The proof that R is left artinian is analogous to the argument in (1). \square

We note in passing that, despite Example 3, every left artinian ring is automatically left noetherian; this result is called the *Hopkins–Levitzky theorem*, proved independently in 1939 by Charles Hopkins and Jacob Levitzki.

A left R -module ${}_R M$ is said to be **simple** if it is nonzero and satisfies the following equivalent conditions:

- (1) The only submodules of M are 0 and M .
- (2) $Rx = M$ for all $0 \neq x \in M$.

¹³¹These are the analogues of the axioms M1–M4 in Section 7.1.

Thus, the simple \mathbb{Z} -modules are the cyclic groups \mathbb{Z}_p where $p \in \mathbb{Z}$ is a prime. Every simple module $_R M$ is principal, that is, $M = Rx$ for some $x \in M$. If $R = D$ is a division ring the converse holds: If $_D M = Dx$ is simple, then $M \cong _D D$ via $dx \leftrightarrow d$. In particular, the simple modules over a field are the one-dimensional vector spaces.

As for groups, a series $M = M_0 \supset M_1 \supset \cdots \supset M_n = 0$ of submodules of a module M is called a **composition series** of length n for M if all the factors M_i/M_{i+1} are simple modules (see Section 9.1).

Theorem 1. *Let $M \neq 0$ be a module.*

- (1) *M has a composition series if and only if it is both noetherian and artinian.*
- (2) **Jordan–Hölder Theorem.** *Any two composition series for M have the same length, and the factors can be paired so that corresponding factors are isomorphic.*

Proof.

(1) If $M = M_0 \supset M_1 \supset \cdots \supset M_n = 0$ is a composition series then M_{n-1} and M_{n-2}/M_{n-1} are both simple, so M_{n-2} is noetherian and artinian by Lemma 2. Then the same is true of M_{n-3} because M_{n-3}/M_{n-2} is simple. Continuing we see that every M_k (including $M_0 = M$) is noetherian and artinian.

Conversely, let M be noetherian and artinian. Since M is artinian, let $K_1 \subseteq M$ be a simple submodule. If $K_1 \neq M$, choose K_2 minimal in the set $\{K \mid K \supset K_1\}$. Then $K_2 \supset K_1 \supset 0$ and K_2/K_1 is simple. If $K_2 \neq M$, let $K_3 \supset K_2 \supset K_1 \supset 0$ where K_3/K_2 simple. Since M is noetherian, this process cannot continue indefinitely, so some $K_n = M$ and we have created a composition series.

(2) The analogue of the proof of the Jordan–Hölder theorem for arbitrary groups (Theorem 1 §9.1) goes through. \blacksquare

The length n of any composition series for M is called the **composition length** of M and denoted $\text{length}(M)$. Note that Lemma 2 and Theorem 1 combine to show that, if $K \subseteq M$ are modules, then M has a composition series if and only if both K and M/K have composition series. Moreover, in this case the proof of Theorem 2 §9.1 goes through to show that

$$\text{length}(M) = \text{length}(K) + \text{length}(M/K), \quad (*)$$

and that the composition factors of M are exactly those of K and M/K .

The finitely generated vector spaces over a field are called *finite dimensional*, and Theorem 6 §6.1 shows that they all have a finite basis. The same is true for any division ring, but we give a different proof using the Jordan–Hölder theorem.

Corollary. *Let D be a division ring and let $_D M$ be a module. Then*

- (1) *M is finitely generated if and only if it has a finite basis.*
- (2) *Any two bases of M contain the same number of elements, say n .*
- (3) *Every nonzero submodule of M has a basis of at most n elements.*

Proof. (1) Since M is finitely generated, let $n \geq 1$ be minimal such that M has a set $\{x_1, \dots, x_n\}$ such that $M = \sum_i D x_i$. If $\sum_i d_i x_i = 0$ and $d_k \neq 0$ for some k , then (since $d_k^{-1} \in D$) it follows that $M = \sum_{i \neq k} D x_i$ contradicting the minimality of n . So $\{x_1, \dots, x_n\}$ is independent and hence is a basis of M . This proves (1).

(2) If $\{x_1, \dots, x_n\}$ is a basis, then $M = Dx_1 \oplus \dots \oplus Dx_n$ where Dx_i is simple, and we obtain a composition series

$$M \supset Dx_2 \oplus \dots \oplus Dx_n \supset \dots \supset Dx_{n-1} \oplus Dx_n \supset Dx_n \supset 0$$

for M . Hence n is the composition length of M , proving (2).

(3) If $K \subseteq {}_D M$ is a submodule, then K has a composition series by Lemma 2 and Theorem 1, and the composition length is at most n by (*). \blacksquare

If D is a division ring, the number of elements in any finite basis of a module ${}_D M$ is called the **dimension** of M and denoted $\dim M$.¹³² With this, most of the theorems about finite dimensional vector spaces (Section 6.1) go through for finitely generated modules over D .

Endomorphism Rings

If R is a ring and M and N are two R -modules, recall that a map $\alpha : M \rightarrow N$ is called **R -linear** (or an **R -morphism**) if $\alpha(x+y) = \alpha(x) + \alpha(y)$ and $\alpha(rx) = r\alpha(x)$ for all $x, y \in M$ and all $r \in R$. Many results about rings (in particular, Wedderburn's theorem) arise from representing them as rings of R -linear maps.

If M and N are two modules write

$$\hom(M, N) = \{\alpha \mid \alpha : M \rightarrow N, \alpha \text{ is } R\text{-linear}\}.$$

If $\alpha, \beta \in \hom(M, N)$, define $\alpha + \beta$ and $-\alpha$ by

$$(\alpha + \beta)(x) = \alpha(x) + \beta(x) \quad \text{and} \quad (-\alpha)(x) = -\alpha(x), \quad \text{for all } x \in M.$$

These are R -linear and make $\hom(M, N)$ into an abelian group. Furthermore, composition of maps distributes over this addition:

$$\begin{aligned} \gamma(\alpha + \beta) &= \gamma\alpha + \gamma\beta \quad \text{whenever } M \xrightarrow{\alpha, \beta} N \xrightarrow{\gamma} K \text{ are } R\text{-linear,} \\ (\alpha + \beta)\delta &= \alpha\delta + \beta\delta \quad \text{whenever } L \xrightarrow{\delta} M \xrightarrow{\alpha, \beta} N \text{ are } R\text{-linear.} \end{aligned}$$

All these routine verifications are left to the reader.

Our interest here is in a special case: If M is any module, an R -linear map $\alpha : M \rightarrow M$ is called an **endomorphism** of M , and we write

$$\text{end } M = \hom(M, M).$$

The additive abelian group $\text{end } M$ becomes a ring, called the **endomorphism ring** of M , if we define addition as above and use composition of maps as the multiplication. Again we leave to the reader the routine verifications of the ring axioms, and that $\text{end } M \cong \text{end } N$ whenever $M \cong N$. Note that the unity of $\text{end } N$ is the identity map 1_M .

If S is a ring, denote the ring of all $n \times n$ matrices over S by $M_n(S)$. Wedderburn's theorem asserts that certain rings are isomorphic to $M_n(D)$ where D is a division ring. The next result shows how matrix rings arise as endomorphism rings. If M is a module, recall that M^n denotes the external direct sum of n copies of M .

Lemma 3. Let ${}_R M$ be a module and write $S = \text{end } M$ for the endomorphism ring. Then $\text{end}(M^n) \cong M_n(S)$ as rings for each integer $n \geq 1$.

¹³²In fact, every nonzero module over a division ring has a basis (and so is free). The proof requires Zorn's lemma (Example 2, Appendix C).

Proof. Identify M^n with the set of n -tuples from M , and let $M \xrightarrow{\sigma_i} M^n \xrightarrow{\pi_i} M$ be the standard maps: $\sigma_i(x) = (0, \dots, x, \dots, 0)$, where x is in position i , and $\pi_i(x_1, x_2, \dots, x_n) = x_i$. Then one verifies that

$$\Sigma \sigma_i \pi_i = 1_{M^n} \quad \text{and} \quad \pi_j \sigma_i = \delta_{ij} 1_M, \quad \text{for all } j \text{ and } i,$$

where $\delta_{ij} = 0$ or 1 according as $i \neq j$ or $i = j$.¹³³ Given $\alpha \in \text{end}(M^n)$, we have $\pi_i \alpha \sigma_j \in S$ for all i and j , so we define

$$\theta : \text{end}(M^n) \rightarrow M_n(S) \quad \text{by} \quad \theta(\alpha) = [\pi_i \alpha \sigma_j].$$

It is routine to see $\theta(\alpha + \beta) = \theta(\alpha) + \theta(\beta)$ for all α and β , and that $\theta(1_{M^n}) = I$, the identity matrix. The (i, j) entry of the matrix $\theta(\alpha)\theta(\beta)$ is

$$\Sigma_k (\pi_i \alpha \sigma_k)(\pi_k \beta \sigma_j) = \pi_i \alpha (\Sigma_k \sigma_k \pi_k) \beta \sigma_j = \pi_i \alpha \beta \sigma_j,$$

so $\theta(\alpha)\theta(\beta) = \theta(\alpha\beta)$. Thus, θ is a ring homomorphism; we claim it is an isomorphism.

If $\theta(\alpha) = 0$ then $\pi_i \alpha \sigma_j = 0$ for all i and j , so

$$\alpha = 1_{M^n} \alpha 1_{M^n} = (\Sigma_i \sigma_i \pi_i) \alpha (\Sigma_j \sigma_j \pi_j) = \Sigma_{i,j} \sigma_i (\pi_i \alpha \sigma_j) \pi_j = 0.$$

This shows that θ is one-to-one, and it remains to show that θ is onto. To this end, let $[\gamma_{ij}] \in M_n(S)$, and define $\alpha : M^n \rightarrow M^n$ by

$$\alpha(x_1, x_2, \dots, x_n) = (\Sigma_t \gamma_{1t}(x_t), \Sigma_t \gamma_{2t}(x_t), \dots, \Sigma_t \gamma_{nt}(x_t)).$$

Then, for every $x \in M$ we have

$$\pi_i \alpha \sigma_j(x) = \pi_i(\alpha(0, \dots, x, \dots, 0)) = \pi_i(\gamma_{1j}(x), \gamma_{2j}(x), \dots, \gamma_{nj}(x)) = \gamma_{ij}(x).$$

It follows that $\pi_i \alpha \sigma_j = \gamma_{ij}$ for all i and j , that is $\theta(\alpha) = [\gamma_{ij}]$. This shows that θ is onto, and so proves the lemma. \blacksquare

If V is an n -dimensional vector space over a field F , then $V \cong F^n$ so Lemma 3 shows that $\text{end}(V) \cong M_n(F)$, a familiar fact from linear algebra.

We say that a mapping $M \xrightarrow{\beta} N$ acts as a *left* operator $x \mapsto \beta(x)$ because we write β on the left of its argument. If $M \xrightarrow{\beta} N \xrightarrow{\alpha} K$, we have the composite map $\alpha\beta : M \rightarrow K$ where $\alpha\beta(x) = \alpha[\beta(x)]$ for all $x \in M$. Hence, because α and β are left operators, the notation $\alpha\beta$ means “first β then α ”. This is somewhat unfortunate because the order gets reversed, but the use of left operators is very common. On the other hand, we could write a map $M \xrightarrow{\beta} N$ on the *right* of its argument, so that $x \mapsto x\beta$. In this case the composite $M \xrightarrow{\beta} N \xrightarrow{\alpha} K$ is given by $x(\beta\alpha) = (x\beta)\alpha$, so we write it as $\beta\alpha : M \rightarrow K$. In particular, $\beta\alpha$ means “first β then α ” in the same order as the arrows. Hence composition means a different thing depending on whether the maps act on the left or the right, and we must always be clear which we are using. Lemma 4 below gives a good illustration of the use of right operators.

An element $e \in R$ is called an **idempotent** if $e^2 = e$. Note that 0 and 1 are idempotents in any ring, and they are the only ones in a division ring (in fact in a domain). If $e \in R$ is an idempotent, we define $eRe \subseteq R$ as follows:

$$eRe = \{ere \mid r \in R\} = \{s \in R \mid es = s = se\} = eR \cap Re.$$

¹³³ δ_{ij} is often called the **Kronecker delta**.

Then eRe is a ring with unity e , called the **corner ring** corresponding to e . The name comes from the following example: If $R = M_2(F) = \begin{bmatrix} F & F \\ F & F \end{bmatrix}$ where F is a field, and if $e = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, then $e^2 = e$ and $eRe = \begin{bmatrix} F & 0 \\ 0 & 0 \end{bmatrix}$.

These corner rings arise as endomorphism rings in a natural way.

Lemma 4. *If R is a ring and $e^2 = e \in R$, then $eRe \cong \text{end}(Re)$ where endomorphisms of Re act as right operators.*

Proof. Given $a \in eRe$, define a right operator

$$\rho_a : Re \rightarrow Re \text{ by } x\rho_a = xa, \text{ for all } x \in Re.$$

Note that $xa \in Re$ because $a \in eRe \subseteq Re$. Then ρ_a preserves addition and, in fact $\rho_a \in \text{end}(Re)$ because $(rx)\rho_a = (rx)a = r(xa) = r(x\rho_a)$ for all $r \in R$ and $x \in Re$.¹³⁴ Hence, we may define

$$\theta : eRe \rightarrow \text{end}(Re) \text{ by } \theta(a) = \rho_a \text{ for each } a \in eRe.$$

We show that θ is a ring isomorphism. We have $\theta(e) = \rho_e = 1_{Re}$, so θ preserves the unity. If $a, b \in eRe$, we have $\rho_{a+b} = \rho_a + \rho_b$ (verify) so θ preserves addition. Also $\rho_{ab} = \rho_a \rho_b$ because $x\rho_{ab} = x(ab) = (xa)b = (x\rho_a)b = x(\rho_a \rho_b)$ for all $x \in Re$, so θ preserves multiplication. Hence θ is a ring homomorphism. Moreover, θ is one-to-one because $\theta(a) = 0$ implies that $\rho_a = 0$, that is, $xa = 0$ for all $x \in Re$. Taking $x = e$ gives that $ea = 0$, so $a = 0$ because $a \in eRe$.

To show that θ is onto, let $\lambda : Re \rightarrow Re$ be R -linear, and define $a = e\lambda$. Then $ae = a$ because $a \in Re$, and $ea = e(e\lambda) = (e^2)\lambda = e\lambda = a$. Hence $a \in eRe$. Finally, if $x \in Re$ then $x = xe$, so $x\lambda = (xe)\lambda = x(e\lambda) = xa = x\rho_a$. Since this holds for all $x \in Re$, we have $\lambda = \rho_a = \theta(a)$, so θ is onto. This completes the proof. ■

Taking $e = 1$ in Lemma 4 gives a very important special case.

Corollary. *If R is a ring then $R \cong \text{end}(RR)$ as rings where the endomorphisms of RR are right multiplications by elements of R .*

Wedderburn's Theorem

Wedderburn's theorem asserts that every simple, left artinian ring is isomorphic to a matrix ring over a division ring. The division ring comes from the next result, due in 1905 to Issai Schur.

Lemma 5. Schur's Lemma. *Let M and N be simple modules.*

- (1) *If $\alpha : M \rightarrow N$ is R -linear, then either $\alpha = 0$ or α is an isomorphism.*
- (2) *$\text{end } M$ is a division ring.*

Proof. (1) Observe that $\ker \alpha$ and $\text{im } \alpha = \alpha(M)$ are submodules of M and N , respectively. If $\alpha \neq 0$ then $\ker \alpha \neq M$ and $\alpha(M) \neq 0$, so $\ker \alpha = 0$ and $\alpha(M) = N$ by simplicity. Hence α is an isomorphism, proving (1). Now (2) follows if $N = M$. ■

In 1907, J.H.M. Wedderburn proved the following important theorem for finite dimensional algebras. A ring R is called simple if the only ideals are 0 and R .

¹³⁴If we wrote ρ_a as a *left* operator then $\rho_a(rx) = (ar)\pi$ and $r\rho_a(x) = (ra)x$ need not be equal.

Theorem 3. Wedderburn's Theorem. The following conditions are equivalent for a ring R :

- (1) R is a simple ring that is left artinian.
- (2) R is a simple ring that has a simple left ideal.
- (3) $R \cong M_n(D)$ for some $n \geq 1$ and some division ring D .
- (4) The right-left analogues of (1) and (2).

Furthermore, the integer n is uniquely determined by R , as is the division ring D up to isomorphism.

Proof. (1) \Rightarrow (2) is clear, and (4) follows by the left-right symmetry in (3).

(2) \Rightarrow (3). Let L be a simple left ideal of R , and recall that LR is the set of all finite sums of products ba where $b \in L$ and $a \in R$. Then LR is an ideal of R containing L , so $LR = R$ because R is a simple ring. In particular, let $1 = b_1a_1 + b_2a_2 + \cdots + b_na_n$ where $n \geq 1$, and where $b_i \in L$ and $a_i \in R$ for each i . Assume that n is the smallest positive integer with this property. The reader can verify that

$$R = La_1 + La_2 + \cdots + La_n. \quad (*)$$

Note that $La_i \neq 0$ for each i because $b_i a_i \neq 0$ by the minimality of n . Hence $L \cong La_i$ because the map $x \mapsto xa_i$ is a nonzero, onto, R -morphism $L \rightarrow La_i$, and L is simple. In particular each La_i is simple.

Now we claim that the sum in (*) is direct. By Theorem 3 §7.1, we must show that $La_k \cap (\sum_{i \neq k} La_i) = 0$ for each k . Suppose not. Then $La_k \cap (\sum_{i \neq k} La_i)$ is a nonzero left ideal contained in La_k . Hence, $La_k \cap (\sum_{i \neq k} La_i) = La_k$ because La_k is simple. But then $La_k \subseteq \sum_{i \neq k} La_i$, and it follows that $R = \sum_{i \neq k} La_i$, contrary to the minimality of n . Hence (*) is a direct sum.

Since $La_i \cong L$ for each i , (*) gives ${}_R R = \bigoplus_{i=1}^n La_i \cong L^n$. But then Lemma 3 and the Corollary to Lemma 4 and give

$$R \cong \text{end}({}_R R) \cong \text{end}(L^n) \cong M_n(\text{end } L).$$

Since $\text{end } L$ is a division ring by Schur's lemma, (3) follows with $D = \text{end } L$.

(3) \Rightarrow (1). The ring $M_n(D)$ is simple by the Corollary to Theorem 7 §3.3 so, given (3), it remains to show that $M_n(D)$ is left artinian. But $M_n(D)$ is a finite dimensional vector space over D (in fact the dimension is n^2), and so is artinian as a D -space. Hence, the ring $M_n(D)$ is left artinian because every left ideal is a D -subspace. This proves (1).

Uniqueness. The fact that ${}_R R \cong L^n$ shows that n is the composition length of ${}_R R$ and so is uniquely determined by R . To show that D is uniquely determined, we prove that $D \cong \text{end } K$ for any simple left ideal K of R . But the proof of (2) \Rightarrow (3) shows that $R \cong K^n$, and hence that $L^n \cong K^n$. We have maps $L \xrightarrow{\sigma_1} L^n \xrightarrow{\tau} K^n \xrightarrow{\pi_i} K$, where τ is an isomorphism, σ_1 is the inclusion, and the π_i are the projections. Then $\pi_i \tau \sigma_1 \neq 0$ for some i because $\tau \sigma_1(L) \neq 0$. Hence $L \cong K$ by Schur's lemma, and so $\text{end } K \cong \text{end } L = D$, as required. ■

Remark. The “left artinian” condition in (1) of Theorem 3 cannot be replaced with “left noetherian”. In fact, there exists a simple, left and right noetherian domain that contains no simple left ideal. It is called the first **Weyl algebra**, after Hermann Weyl, and can be described roughly as the ring of polynomials over \mathbb{R} in noncommuting indeterminates x and y which satisfy the condition that $xy - yx = 1$.

This ring first arose in quantum mechanics as an algebra generated by position and momentum operators.

Exercises 11.1

1. Show that a module $_RM$ is simple if and only if $M \cong R/L$ for some maximal left ideal L .
2. Show that the following are equivalent for a ring R : (1) R is a division ring; (2) every principal module $Rx \neq 0$ is simple; (3) $_RR$ is simple.
3. If $aR = bR$ where $a, b \in R$, show that there is an R -isomorphism $\sigma : Ra \rightarrow Rb$ such that $\sigma(a) = b$.
4. Given $_RM$ and $m \in M$ define $\lambda_m : R \rightarrow M$ by $\lambda_m(a) = am$ for all $a \in R$.
 - (a) Show that λ_m is R -linear for each $m \in M$.
 - (b) Show that $m \mapsto \lambda_m$ is an abelian group isomorphism $M \rightarrow \text{hom}(_RR, M)$.
5. If R is left noetherian (left artinian), show that the same is true of the corner ring eRe for any idempotent $e = e^2$ in R . [Hint: If $L \subseteq eRe$ is a left ideal, consider RL .]
6. Show that the following are equivalent for a ring R : (1) R is left noetherian (left artinian); (2) $M_n(R)$ is left noetherian (left artinian). [Hint: $M_n(R)$ is a free left R -module.]
7. If R is left noetherian, show that every finitely generated left R -module is left noetherian. [Hint: Lemma 2 and Theorem 5 §7.1.]
8. Let R be left artinian. If X is a subset of a left module $_RM$, define the annihilator $\text{ann}(X) = \{a \in R \mid ax = 0 \text{ for all } x \in X\}$.
 - (a) Show that $\text{ann}(M) = \text{ann}(X)$ for some finite subset $X \subseteq M$.
 - (b) If $\text{ann}(M) = 0$, show that $_RR$ is isomorphic to a submodule of M .
9. Complete the solution of Example 2 as follows: If $p \in \mathbb{Z}$ is a prime and we write $X = \{\frac{m}{p^k} \in \mathbb{Q} \mid m \in \mathbb{Z}, k \geq 0\}$, show that the only subgroups of X that contain \mathbb{Z} are $\mathbb{Z}\frac{1}{p^k}$ for $k \geq 0$. [Hint: If $\mathbb{Z} \subset Y \subset X$, Y a subgroup of X , choose $\frac{m}{p^n}$ in Y where m and p^n are relatively prime and n is maximal.]
10. Let $K \subseteq M$ be modules. If $K \subseteq N \subseteq M$ where N is a submodule, show that N/K is a submodule of M/K , and that every submodule \mathcal{X} of M/K has the form $\mathcal{X} = N/K$ for some submodule $N \supseteq K$. [Hint: Theorem 5 § 8.1.]
11. Let $_RM$ be a module and let $\pi^2 = \pi \in \text{end } M$.
 - (a) Show that $M = \pi(M) \oplus \ker \pi$.
 - (b) If $M = N \oplus K$, N and K submodules, show that $\pi^2 = \pi \in \text{end } M$ exists such that $N = \pi(M)$ and $K = \ker \pi$.
 - (c) Show that $\pi(M) = \ker(1 - \pi)$ and $(1 - \pi)(M) = \ker \pi$.
12. Show that a module M is noetherian if and only if every submodule is finitely generated.
13. Call a module M **finite dimensional** if it contains no infinite direct sum of nonzero submodules, and call M **indecomposable** if it is not a direct sum $M = A \oplus B$ where A and B are both nonzero submodules.
 - (a) Show that M is finite dimensional if it is either noetherian or artinian.
 - (b) If M is finite dimensional show that $M = N_1 \oplus N_2 \oplus \cdots \oplus N_k$ where each N_j is indecomposable.
14. Suppose R is a ring for which $_RR$ is finite dimensional (preceding exercise). If $ab = 1$ in R show that $ba = 1$. [Hint: Write $ba = e$ and show that $e^2 = e$ and $_RR \cong Ra = Re$.]

15. Show that the converse of (2) of Schur's lemma is not true. That is find a module K such that $\text{end } K$ is a division ring but K is not simple. [Hint: If F is a field, consider $R = \begin{bmatrix} F & F \\ 0 & F \end{bmatrix}$, let $e = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, and consider Re .]
16. Extend Lemma 4 as follows: If $e^2 = e$ and $f^2 = f$ are idempotents in a ring R , then $eRf \cong \text{hom}(Re, Rf)$ as additive abelian groups. [Hint: If $a \in eRf$ and $x \in Re$ then $xa \in Rf$.]
17. Let M be a module and let $\alpha \in \text{end}(M)$.
- If M is noetherian and α is onto, show that α is one-to-one.
 - If M is artinian and α is one-to-one, show that α is onto.
- [Hint: We have chains $\ker(\alpha) \subseteq \ker(\alpha^2) \subseteq \dots$ and $\alpha(M) \supseteq \alpha^2(M) \supseteq \dots$.]
18. **Fittings Lemma.** Suppose that the module M is both artinian and noetherian. If $\alpha \in \text{end}(M)$ show that there exists $n \geq 1$ such that $M = \alpha^n(M) \oplus \ker(\alpha^n)$. [Hint: $\ker(\alpha) \subseteq \ker(\alpha^2) \subseteq \dots$ and $\alpha(M) \supseteq \alpha^2(M) \supseteq \dots$.]
19. Let R be an n -dimensional algebra over a field F , and fix a basis $\{u_1, u_2, \dots, u_n\}$ of R . Define $\theta : R \rightarrow M_n(F)$ by $\theta(r) = [r_{ij}]$, where $u_i r = \sum_{j=1}^n r_{ij} u_j$.
- Show that θ is a one-to-one ring homomorphism.
- The image $\theta(R)$ is called the **regular representation** of R ; in each case find it.
- $R = \mathbb{C}$, $F = \mathbb{R}$, basis $\{1, i\}$.
 - $R = \mathbb{H}$, $F = \mathbb{R}$, basis $\{1, i, j, k\}$.
 - Basis $\{1, u\}$ where $1^2 = 1$, $1u = u = u1$ and $u^2 = 0$.
 - Basis $\{1, u\}$ where $1^2 = 1$, $1u = u = u1$ and $u^2 = 1$.

11.2 THE WEDDERBURN–ARTIN THEOREM

The proof of Wedderburn's theorem (Theorem 3 §11.1) shows that if R is a simple ring containing a simple left ideal L , then $R = L_1 \oplus L_2 \oplus \dots \oplus L_n$ where the L_i are isomorphic, simple left ideals (in fact $L_i \cong L$ for each i). Wedderburn actually treated the case where the L_i are *not* necessarily all isomorphic, albeit in the special case when R is a finite dimensional algebra. We deal with this general case in this section, in the context of rings.

The work necessitates a look at modules that are the sum of a (possibly infinite) family of simple submodules. This investigation provides an extension of the well-known theory of vector spaces over any division ring. However, not surprisingly, transfinite methods are required.

The technique we need is called Zorn's lemma. Let \mathcal{S} be a nonempty family of subsets of some set. A **chain** from \mathcal{S} is a (possibly infinite) set $\{X_i \mid i \in I\} \subseteq \mathcal{S}$ where either $X_i \subseteq X_j$ or $X_j \subseteq X_i$ for any $i, j \in I$. We say that \mathcal{S} is **inductive** if the union of any chain from \mathcal{S} is again in \mathcal{S} . **Zorn's lemma** asserts that every inductive family \mathcal{S} has a **maximal** member, that is a set $Y \in \mathcal{S}$ such that $Y \subseteq Z \in \mathcal{S}$ implies $Y = Z$. A more general version of Zorn's lemma is discussed in Appendix C.

Semisimple Modules

Let R be a ring, and let $\{M_i \mid i \in I\}$ be a (possibly infinite) family of submodules of a module M . The **sum** $\sum_{i \in I} M_i$ of these submodules is defined to be the set of

all finite sums of elements of the M_i ; more formally

$$\Sigma_{i \in I} M_i = \{x_1 + \cdots + x_m \mid m \geq 1, x_j \in M_j \text{ for some } j \in I\}.$$

This is a submodule of M that contains every M_i . The following lemma is the extension of Theorem 3 §7.1 to the case of infinite sets of submodules.

Lemma 1. *The following are equivalent for submodules $\{M_i \mid i \in I\}$ of a module:*

- (1) *Every element of $\Sigma_{i \in I} M_i$ is uniquely represented as a sum of elements of M_i for distinct i .*
- (2) *The only way a sum of elements from distinct M_i can equal 0 is if each of the elements is zero.*
- (3) $\Sigma_{i \in J} M_i$ is direct for every finite subset $J \subseteq I$.

Proof. (1) \Rightarrow (2). If $x_1 + \cdots + x_m = 0 = 0 + \cdots + 0$ then each $x_i = 0$ by (1).

(2) \Rightarrow (3). If $N = M_{i_1} + \cdots + M_{i_m}$ where the i_j are distinct, then N is direct by (2) and Theorem 3 §7.1.

(3) \Rightarrow (1). If $x \in M$ let $x = x_1 + \cdots + x_m = y_1 + \cdots + y_k$ where $x_j \in M_{i_j}$ and $y_k \in M_{i_k}$. Inserting zeros where necessary, we may assume that x_i and y_i are in M_{i_j} for each i , and that these M_{i_j} are distinct. Hence, $x_1 + \cdots + x_m = y_1 + \cdots + y_k$ in $M_{i_1} \oplus \cdots \oplus M_{i_n}$ for distinct i_j by (3). Thus, $x_i = y_i$ for each i by Theorem 3 §7.1, proving (1). \blacksquare

In this case we say that $\Sigma_{i \in I} M_i$ is a **direct sum**, and we write it as $\oplus_{i \in I} M_i$. One reason for introducing these infinite direct sums here is that they provides the language needed to generalize some of the results about vector spaces.

If D is a division ring and ${}_D M$ is an module, let $B = \{b_i \mid i \in I\}$ be a (possibly infinite) set of nonzero elements of M . Then B is said to **span** M if $M = \Sigma_{i \in I} Db_i$. The set B is called **independent** if $\Sigma r_i b_i = 0$, $r_i \in D$, implies that $r_i = 0$ for each i , equivalently if $M = \oplus_{i \in I} Db_i$ (since D is a division ring). A set B that is both independent and spans M is called a **basis** of M . Furthermore, each principal module Db_i is simple (again because D is a division ring). In this form, these facts can be stated much more generally.

It is shown in Appendix C that every module M over a division ring D has a basis. By the above discussion this means that M is a direct sum of simple submodules. In general, a module ${}_R M$ over a ring R is called **semisimple** if it is the direct sum of a (possibly infinite) family of simple submodules. The following example shows that ${}_R R$ is semisimple if R is the 2×2 matrix ring over a division ring.

Example 1. Let D be a division ring, and let $R = M_2(D) = \begin{bmatrix} D & D \\ D & D \end{bmatrix}$. If $L_1 = \begin{bmatrix} D & 0 \\ D & 0 \end{bmatrix}$ and $L_2 = \begin{bmatrix} 0 & D \\ 0 & D \end{bmatrix}$, show that L_1 and L_2 are each simple left ideals of R , that $R = L_1 \oplus L_2$, and that $L_1 \cong L_2$ as R -modules.

Solution. L_1 and L_2 are left ideals by the definition of matrix multiplication, and it is clear that $R = L_1 + L_2$ and $L_1 \cap L_2 = 0$. Hence $R = L_1 \oplus L_2$. We show that L_1 is simple; the proof for L_2 is similar. So let $0 \neq x = \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} \in L_1$, say $a \neq 0$.

Given $\begin{bmatrix} r & 0 \\ s & 0 \end{bmatrix} \in L_1$ we have $\begin{bmatrix} r & 0 \\ s & 0 \end{bmatrix} = \begin{bmatrix} ra^{-1} & 0 \\ sa^{-1} & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix} \in L_1$, so $L_1 = Rx$.

A similar argument shows that $L_1 = Rx$ when $b \neq 0$.

Finally, if $a = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ then it is a routine matter to show that the maps $L_1 \rightarrow L_2$ given by $x \mapsto xa$ and $L_2 \rightarrow L_1$ given by $x \mapsto xb$ are mutually inverse R -isomorphisms. Hence $L_1 \cong L_2$. \square

We are going to give several characterizations of semisimple modules, and the following notion is essential. A module M is said to be **complemented** if every submodule K is a direct summand, that is if $M = K \oplus N$ for some submodule N . Hence every simple module is complemented, as is every finite dimensional vector space (Theorem 8(3) §6.1). A submodule N of a module M is called **proper** if $N \neq M$, and a maximal proper submodule is called simply a **maximal** submodule.

Lemma 2. Let M be a complemented module. Then

- (1) Every submodule N of M is complemented.
- (2) Every proper submodule of M is contained in a maximal submodule.

Proof. (1) If $K \subseteq N$ let $K \oplus K_1 = M$. Then $N = N \cap (K \oplus K_1) = K \oplus (N \cap K_1)$ by the modular law (Theorem 2 §7.1).

(2) If $x \notin N$, write $\mathcal{S} = \{P \mid P \text{ is a submodule, } N \subseteq P \text{ and } x \notin P\}$. Then \mathcal{S} is nonempty and inductive (verify) so, by Zorn's lemma, choose a submodule P maximal in \mathcal{S} . Since M is complemented, let $M = P \oplus Q$; we show that P is maximal by showing that Q is simple. If not, let $A \subseteq Q$ where $A \neq 0, Q$. By (1) write $Q = A \oplus B$. Then $P \oplus A$ and $P \oplus B$ both strictly contain P , so neither is in \mathcal{S} . Hence x lies in both by the maximality of P , say $x = p_1 + a = p_2 + b$ where $p_1, p_2 \in P$, $a \in A$ and $b \in B$. Since $a \in Q$ and $b \in Q$, and since $P \oplus Q$ is direct, it follows that $a = b$. But then $a = b \in A \cap B = 0$, so $x = p_1 \in P$, a contradiction. \blacksquare

Lemma 3. Let $M = \sum_{i \in I} M_i$ where each M_i is simple. If N is any submodule of M there exists $J \subseteq I$ such that $M = N \oplus (\bigoplus_{j \in J} M_j)$. In particular, M is complemented.

Proof. If $N = M$ take $J = \emptyset$. If $N \neq M$ then some $M_j \not\subseteq N$ so $N \cap M_j = 0$ (because M_j is simple). Hence, \mathcal{S} is nonempty where

$$\mathcal{S} = \{J \subseteq I \mid N + \sum_{i \in J} M_i \text{ is a direct sum}\}.$$

Let $\{J_i\}$ be a chain from \mathcal{S} , and write $J = \cup J_i$. To see that J is in \mathcal{S} , let

$$0 = x + x_{i_1} + \cdots + x_{i_n} \in N + \sum_{i \in J} M_i,$$

where $x \in N$ and each $x_{i_t} \in M_{i_t}$. Since $\{J_i\}$ is a chain, there exists some J_k containing all these i_t so (since $J_k \in \mathcal{S}$) it follows that $x = x_{i_t} = 0$ for each t . Hence $N + \sum_{i \in J} M_i$ is direct, which shows that $J \in \mathcal{S}$.

Hence, by Zorn's lemma, \mathcal{S} has a maximal member J_0 . If $L = N + \sum_{i \in J_0} M_i$, it remains to show that $L = M$. Since $M = \sum M_i$, we show that $M_i \subseteq L$ for each $i \in I$. This is clear for all $i \in J_0$. If $i_0 \in I - J_0$, then $J_0 \cup \{i_0\} \notin \mathcal{S}$ so $L + M_{i_0}$ is not direct, that is, $L \cap M_{i_0} \neq 0$. Hence, $L \cap M_{i_0} = M_{i_0}$ as M_{i_0} is simple, so $M_{i_0} \subseteq L$, as required. \blacksquare

With this we can characterize the semisimple modules.

Theorem 1. The following conditions are equivalent for a module $M \neq 0$:

- (1) $M = \sum_{i \in I} M_i$, where each M_i is a simple submodule.

- (2) $M = \bigoplus_{i \in I} M_i$, where each M_i is a simple submodule.
- (3) M is complemented.
- (4) Every maximal submodule of M is a direct summand, and every proper submodule is contained in a maximal submodule.

Proof. (1) \Rightarrow (2) and (2) \Rightarrow (3). These are by Lemma 3, the first with $N = 0$.

(3) \Rightarrow (4). This is clear by Lemma 2.

(4) \Rightarrow (1). Let S denote the sum of all simple submodules of M (take $S = 0$ if there are none). Suppose $S \neq M$. By (4) let $S \subseteq N$ where N is maximal in M , and write $M = N \oplus K$ for some submodule K , again by (4). Then K is simple (because $K \cong M/N$) so $K \subseteq S \subseteq N$, contrary to $K \cap N = 0$. So $S = M$ and (1) follows. ■

A module $M \neq 0$ is called **semisimple** if it satisfies the conditions in Theorem 1. We also regard 0 as a semisimple module. Hence every module over a division ring is semisimple, and the ring in Example 1 is semisimple as a left module over itself. Lemma 3 gives

Corollary 1. If M is semisimple so also is every submodule and image of M . More precisely, if $N \subseteq M = \bigoplus_{i \in I} M_i$ where each M_i is simple, then

$$N \cong \bigoplus_{i \in J} M_i \quad \text{and} \quad M/N \cong \bigoplus_{i \in I-J} M_i \quad \text{for some } J \subseteq I.$$

Note that we need not have $N = \bigoplus_{i \in J} M_i$ in Corollary 1. As an example, let $M = F^2 = F \oplus F$ where F is a field. Then $M = M_1 \oplus M_2$ where $M_1 = F \oplus 0$ and $M_2 = 0 \oplus F$. But if $N = \{(a, a) \mid a \in F\}$ then $M_i \not\subseteq N$ for $i = 1, 2$.

Corollary 2. If M is finitely generated then M is semisimple if and only if every maximal submodule is a direct summand.

Proof. By Theorem 1, it remains to show that every proper submodule $K \subset M$ is contained in a maximal submodule of M . To this end, let $\mathcal{S} = \{{}_R X \mid K \subseteq X \neq M\}$; we show that \mathcal{S} contains maximal members (they are then maximal in M). So let $\{X_i \mid i \in I\}$ be a chain from \mathcal{S} and write $X = \cup_{i \in I} X_i$. By Zorn's lemma it suffices to show that $X \neq M$. But if $X = M$, let $M = Rm_1 + \cdots + Rm_k$ (as M is finitely generated). Then each $m_t \in X_{i_t}$ for some $i_t \in I$. But the X_i are a chain so there exists $k \in I$ such that $X_{i_t} \subseteq X_k$ for each i_t . Hence each $m_t \in X_k$, so $M \subseteq X_k$, a contradiction. Hence $X \neq M$ after all. ■

Lemma 4. The following are equivalent for a semisimple module M :

- (1) M is finitely generated.
- (2) $M = M_1 \oplus M_2 \oplus \cdots \oplus M_n$ where each M_i is simple.
- (3) M is artinian.
- (4) M is noetherian.

Finally the decomposition in (2) is unique: If $M = N_1 \oplus N_2 \oplus \cdots \oplus N_m$, where each N_i is simple then $m = n$ and (after relabeling) $N_i \cong M_i$ for each i .

Proof. (1) \Rightarrow (2). If $M = \bigoplus_i M_i$ where the M_i are simple, the generators of M all lie in a finite sum $M_{i_1} \oplus M_{i_2} \oplus \cdots \oplus M_{i_n}$ by (1).

(2) \Rightarrow (3) and (2) \Rightarrow (4). By (2), $M \supseteq M_2 \oplus \cdots \oplus M_n \supseteq \cdots \supseteq M_n \supseteq 0$ is a composition series for M so (3) and (4) follow from Theorem 1 §11.1.

(3) \Rightarrow (1) and (4) \Rightarrow (1). If $M = \bigoplus_{i \in I} M_i$ where I is infinite, we may assume that $\{1, 2, 3, \dots\} \subseteq I$. Then $(M_1 \oplus M_2 \oplus M_3 \oplus \dots) \supset (M_2 \oplus M_3 \oplus \dots) \supset \dots$ contradicts (3) and $M_1 \subset M_1 \oplus M_2 \subset \dots$ contradicts (4).

Finally, as we saw above, (2) gives rise to a composition series M with factors M_1, M_2, \dots, M_n . Hence, the last statement in the corollary follows from the Jordan–Hölder theorem (Theorem 1 §11.1). \blacksquare

Note that Lemma 4 proves the uniqueness of the number of elements in a finite basis of a module over any division ring (so we can speak of dimension). In fact, a version of the uniqueness actually goes through for *arbitrary* infinite direct sum decompositions of a semisimple module, a result beyond the scope of this book.

Homogeneous Components

Before proceeding we need a technical lemma.

Lemma 5. *Let $M = \bigoplus_{i \in I} M_i$ with each M_i simple and let $K \subseteq M$ be a simple submodule. Then there exist i_1, i_2, \dots, i_m in I such that $K \subseteq M_{i_1} \oplus M_{i_2} \oplus \dots \oplus M_{i_m}$ and $K \cong M_{i_t}$ for each $t = 1, 2, \dots, m$.*

Proof. Choose $0 \neq y \in K$, and write $y = y_{i_1} + y_{i_2} + \dots + y_{i_m}$, where $0 \neq y_{i_t} \in M_{i_t}$ for each t . Then $K \cap (M_{i_1} \oplus M_{i_2} \oplus \dots \oplus M_{i_m}) \neq 0$ so, because K is a simple module, $K \subseteq M_{i_1} \oplus M_{i_2} \oplus \dots \oplus M_{i_m}$. Because this is a direct sum, we can define $\alpha_t : K \rightarrow M_{i_t}$ as follows: If $x \in K$ take $\alpha_t(x) = x_{i_t}$, where $x = x_{i_1} + x_{i_2} + \dots + x_{i_m}$ and $x_{i_j} \in M_{i_j}$ for each j . Then α_t is R -linear for each t , and $\alpha_t \neq 0$ because $\alpha_t(y) = y_{i_t} \neq 0$. So α_t is an isomorphism by Schur's lemma. \blacksquare

If $K \subseteq M$ are modules with K simple, define the **homogeneous component** $H(K)$ of M generated by K as follows:

$$H(K) = \Sigma\{X \subseteq M \mid X \text{ is a submodule and } X \cong K\}.$$

We say that $H(K) = 0$ if M contains no copy of K . Hence, if $K \subseteq M$ is simple then $H(K)$ is a submodule of M containing every submodule isomorphic to K (and hence K itself). In fact every simple submodule of $H(K)$ is isomorphic to K .

Lemma 6. *Let $K \subseteq M$ be modules with K simple. The following are equivalent for a simple submodule $L \subseteq M$:*

- (1) $L \cong K$.
- (2) $H(L) = H(K)$.
- (3) $L \subseteq H(K)$.

Proof. (1) \Rightarrow (2). This is because $X \cong L$ if and only $X \cong K$ by (1).

(2) \Rightarrow (3). Here $L \subseteq H(L) = H(K)$ by (2).

(3) \Rightarrow (1). If $L \subseteq H(K) = \Sigma\{X \subseteq M \mid X \cong K\}$, then $L \cong X$ for some $X \cong K$ by Lemma 5. This proves (1). \blacksquare

A submodule N of M is said to be **fully invariant** in M if $\alpha(N) \subseteq N$ for every endomorphism $\alpha : M \rightarrow M$. Clearly 0 and M are fully invariant for every module M . Recall that the endomorphisms of ${}_R R$ are just the right multiplications by elements of R (Corollary to Lemma 4 §11.1). It follows that a left ideal L of R is fully invariant in ${}_R R$ if and only if L is an ideal of R .

We can now prove the main structure theorem for semisimple modules. For convenience, we say that submodules K and N meet if $K \cap N \neq 0$.

Theorem 2. *Let M be a semisimple module. Then the following hold:*

- (1) *M is the direct sum of its homogeneous components.*
- (2) *Each homogeneous component of M is fully invariant in M .*
- (3) *Each fully invariant submodule of M is the direct sum of the homogeneous components it meets.*

Proof. Let $\{H(K_i) \mid i \in I\}$ be the distinct homogeneous components of M .

(1) $M = \sum_{i \in I} H(K_i)$ because M is semisimple; to see that this sum is direct we show that $H(K_t) \cap [\sum_{i \neq t} H(K_i)] = 0$ for each $t \in I$, and invoke Lemma 1. But if $H(K_t) \cap [\sum_{i \neq t} H(K_i)] \neq 0$ then it contains a simple submodule L (being semisimple). Then $L \cong K_t$ by Lemma 5 because $L \subseteq H(K_t)$; and $L \cong K_i$ for some $i \neq t$ by Lemma 5 because $L \subseteq \sum_{i \neq t} H(K_i)$. Hence, $K_t \cong L \cong K_i$ and so $H(K_t) = H(K_i)$ by Lemma 6, contrary to the choice of the $H(K_i)$.

(2) Consider $H(K)$, where $K \subseteq M$ is simple. If $\alpha \in \text{end } M$ we must show that $\alpha[H(K)] \subseteq H(K)$, that is $\alpha(L) \subseteq H(K)$ whenever $L \subseteq M$ and $L \cong K$. But since L is simple, either $\alpha(L) = 0$ or $\alpha(L) \cong L \cong K$ by Schur's lemma. Either way, $\alpha(L) \subseteq H(K)$.

(3) Let the K_i be as above, and let $N \subseteq M$ be fully invariant. We show that $N = \sum \{H(K_i) \mid H(K_i) \cap N \neq 0\}$. Every simple submodule of M is in some $H(K_i)$ by Lemma 5, so $N \subseteq \sum \{H(K_i) \mid H(K_i) \cap N \neq 0\}$ because N is semisimple. Conversely, if $H(K_i) \cap N \neq 0$ then there exists $U \subseteq H(K_i) \cap N$ such that $U \cong K_i$. If $L \cong K_i$ is arbitrary then $L \cong U$ by Lemma 5, so let $\sigma : U \rightarrow L$ be an isomorphism. Since M is complemented (Theorem 1), write $M = U \oplus U_1$, and define $\alpha : M \rightarrow M$ by $\alpha(u + u_1) = \sigma(u)$. Then $L = \sigma(U) = \alpha(U) \subseteq \alpha(N) \subseteq N$, where $\alpha(N) \subseteq N$ because N is fully invariant. Since $L \cong K_i$ was arbitrary, it follows that $H(K_i) \subseteq N$. ■

A semisimple module M is called **homogeneous** if it has only one homogeneous component, that is (by Lemma 6) if all simple submodules of M are isomorphic. In particular if D is a division ring and $R = M_2(D)$, then Example 1 shows that ${}_R R$ is a homogeneous, semisimple module. In fact much more is true as we shall see below.

Free and Projective Modules

The Wedderburn–Artin theorem shows that rings R such that ${}_R R$ is semisimple as a left R -module are isomorphic to finite direct products of matrix rings over division rings. Other equivalent conditions on R are also considered.

Finitely generated free and projective modules were discussed in Section 7.1. However, the Wedderburn–Artin theorem involves arbitrary projective modules, so we pause to review these notions. Let ${}_R W$ be a module, and let B be a set of nonzero elements of W . We make three definitions:

B is said to generate W if $W = \sum_{w \in B} R w$.

B is called independent if $\sum_i r_i w_i = 0$, $r_i \in R$, $w_i \in B$, implies each $r_i = 0$.

B called a basis of W if it is independent and generates W .

A module ${}_R W$ that has a basis is called a **free module**. Note that the second condition above implies that $R w_i \cong {}_R R$ for each $i \in B$ (see the corollary to Theorem

1 §7.1). Hence, every free module is isomorphic to a direct sum of copies of R . The finitely generated free modules in Section 7.1 are examples, but free modules with bases of arbitrary size can easily be constructed.

If I is any nonempty set and R is any ring, there exists a free R -module with a basis indexed by I . If $i \mapsto r_i$ is any function $I \rightarrow R$, write it as $\langle r_i \rangle$. Hence $\langle r_i \rangle = \langle s_i \rangle$ if and only if $r_i = s_i$ for each $i \in I$, and we call r_i the *i*th component of $\langle r_i \rangle$. We call $\langle r_i \rangle$ an *I-sequence* from R . Define

$$R^{(I)} = \{ \langle r_i \rangle \mid r_i = 0 \text{ for all but finitely many } i \in I \}.$$

If I has n elements it is clear that $R^{(I)} = R^n$. In general, $R^{(I)}$ becomes a left R -module with componentwise operations:

$$\langle r_i \rangle + \langle s_i \rangle = \langle r_i + s_i \rangle \quad \text{and} \quad r \langle r_i \rangle = \langle rr_i \rangle, \quad \text{for all } r \in R.$$

If e_i denotes the *I*-sequence with *i*th component 1 and all other components 0, then it is a routine matter to check that $\{e_i \mid i \in I\}$ is a basis of $R^{(I)}$, called the **standard basis**. Hence, $R^{(I)} = \bigoplus_{i \in I} Re_i$, where $Re_i \cong_R R$ for each i .

Now let $\{x_i \mid i \in I\}$ be a generating set for a module $_RM$, that is, $M = \sum_{i \in I} Rx_i$. Let W be a free module with a basis $B = \{w_i \mid i \in I\}$ indexed by I (for example $W = R^{(I)}$). Given r_1, r_2, \dots, r_n in R , define

$$\beta : W \rightarrow M \quad \text{by} \quad \beta(r_1w_1 + r_2w_2 + \cdots + r_nw_n) = r_1x_1 + r_2x_2 + \cdots + r_nx_n.$$

Because B is a basis, this map is well defined, and it is evidently onto. Since every module M has a generating set (the set of all nonzero elements, for example), this proves the first part of

Lemma 7. *Let M denote any left R -module.*

- (1) *M is an image of a free module.*
- (2) *If $\alpha : M \rightarrow W$ is onto and W is free, then $\ker \alpha$ is a direct summand of M .*

Proof. We proved (1) above. As to (2), let $B = \{w_i \mid i \in I\}$ be a basis of W . As α is onto, let $w_i = \alpha(x_i)$, $x_i \in M$ for each i . By the above discussion, there exists $\beta : W \rightarrow M$ such that $\beta(w_i) = x_i$ for each i . Then $\alpha\beta(w_i) = \alpha(x_i) = w_i$ for each i , so $\alpha\beta = 1_W$ because the w_i generate W . But then $M = \ker \alpha \oplus \beta(W)$ by Lemma 8 below, proving (2). \blacksquare

Lemma 8. *If $\alpha : M \rightarrow P$ is onto, the following conditions are equivalent:*

- (1) *There exists $\beta : P \rightarrow M$ such that $\alpha\beta = 1_P$.*
- (2) *$\ker \alpha$ is a direct summand of M , in fact $M = \ker \alpha \oplus \beta(P)$.*

In this case the map α is said to split.

Proof. (1) \Rightarrow (2). If $\alpha\beta = 1_P$ as in (1), let $m \in M$. As $\alpha(m) \in P$, we have $\alpha(m) = 1_P[\alpha(M)] = \alpha\beta\alpha(m)$. Thus $m - \beta\alpha(m) \in \ker \alpha$, so $M = \ker \alpha + \beta(P)$. But if $m \in \ker \alpha \cap \beta(P)$, let $m = \beta(p)$, $p \in P$. Then $0 = \alpha(m) = \alpha\beta(p) = p$, so $m = \beta(p) = \beta(0) = 0$. This proves that $\ker \alpha \cap \beta(P) = 0$ and so proves (2).

(2) \Rightarrow (1). Given (2), let $M = \ker \alpha \oplus Q$. Observe that $P = \alpha(M) = \alpha(Q)$, so we define $\beta : P \rightarrow M$ as follows: if $p \in P$ and $p = \alpha(q)$, $q \in Q$, define $\beta(p) = q$. This is well defined because if $p = \alpha(q_1)$, $q_1 \in Q$, then $q - q_1 \in Q \cap \ker \alpha = 0$. Now given $p = \alpha(q)$ in P then $\alpha\beta(p) = \alpha(q) = p$, proving (1). \square

A module $R P$ is called **projective** if it satisfies the condition:

If $R M \xrightarrow{\alpha} P$ is R -linear and onto then $\ker \alpha$ is a direct summand of M .

Hence all free modules are projective by Lemma 7 (but see Example 3 below). Also, P is projective if and only if every onto R -morphism $M \rightarrow P$ splits. We need

Lemma 9. If P is projective and $Q \cong P$, then Q is projective.

Proof. Let $\alpha : M \rightarrow Q$ be onto. If $\sigma : Q \rightarrow P$ is an isomorphism then $\sigma\alpha : M \rightarrow P$ is onto so $\ker \alpha = \ker \sigma\alpha$ is a direct summand of M . Hence, Q is projective. \square

Theorem 3. The following conditions on a module $R P$ are equivalent:

- (1) P is projective.
- (2) P is isomorphic to a direct summand of a free module.
- (3) If α is onto in the diagram, then γ exists such that $\alpha\gamma = \beta$.
- (4) If $M \xrightarrow{\alpha} P$ is onto there exists $P \xrightarrow{\beta} M$ such that $\alpha\beta = 1_P$.

$$\begin{array}{ccc} & P & \\ \gamma \swarrow & & \downarrow \beta \\ M & \xrightarrow{\alpha} & N \end{array}$$

Proof. (1) \Rightarrow (2). If $W \xrightarrow{\alpha} P$ is onto, then $W = \ker \alpha \oplus Q$ for some Q by (1). Hence, $P = \alpha(P) \cong W/\ker \alpha \cong Q$, proving (2).

(2) \Rightarrow (3). Since being projective is preserved under isomorphism, we may assume that $W = P \oplus Q$ is free. Let $\pi : W \rightarrow P$ be the projection defined by $\pi(p+q) = p$ for $p \in P$ and $q \in Q$, and let $\{w_i \mid i \in I\}$ be a basis of W . Given α as in the diagram, we must construct $\gamma : P \rightarrow M$ such that $\alpha\gamma = \beta$.

Note that $\beta\pi(w_i) \in N$ for each $i \in I$. Since $\alpha : M \rightarrow N$ is onto, there exists $m_i \in M$ such that $\beta\pi(w_i) = \alpha(m_i)$. But the w_i are a basis of W , so there exists $\lambda : W \rightarrow M$ such that $\lambda(w_i) = m_i$ for each i . Hence,

$$\alpha\lambda(w_i) = \alpha(m_i) = \beta\pi(w_i), \quad \text{for each } i.$$

It follows that $\alpha\lambda = \beta\pi$ because the w_i generate W . Finally, let $\gamma : P \rightarrow M$ be the restriction of λ to P , that is, $\gamma(p) = \lambda(p)$ for all $p \in P$. Compute

$$\alpha\gamma(p) = \alpha\lambda(p) = \beta\pi(p) = \beta(p), \quad \text{for all } p \in P,$$

because $\pi(p) = p$. Hence, $\alpha\gamma = \beta$, as required.

(3) \Rightarrow (4). Take $N = P$ and $\beta = 1_P$ in the diagram.

(4) \Rightarrow (1). This is Lemma 8. \square

Example 2. A module $R P$ is a principal projective module if and only if $P \cong Re$ for some $e^2 = e \in R$.

Solution. If $P = Rx$ is projective, define $\alpha : R \rightarrow Rx$ by $\alpha(r) = rx$ for all $r \in R$. Then α is onto so $R R = \ker \alpha \oplus L$ for some left ideal L . Hence $P \cong R/\ker \alpha \cong L$. But direct summands of $R R$ have the form Re for some idempotent e by Example 6 §7.1, so we have $P \cong Re$ for some $e^2 = e \in R$.

Conversely, if $e^2 = e$ then $Re \oplus R(1 - e) = R$ is free, so Re is projective by Theorem 3. \square

Example 3. There exist projective modules that are not free.

Solution. Let F be a field, and let $R = F \times F$. Then $e = (1, 0)$ is an idempotent in R , so $Re = \{(a, 0) \mid a \in F\}$ is projective. But $\dim_F(Re) = 1$ so Re is not free since any free R -module has F -dimension at least 2 (because $\dim_F R = 2$). \square

The Wedderburn–Artin Theorem

Let A be a left ideal of a ring R . If X is a nonempty subset of a module $_RM$, define AX to be the set of all finite sums of elements ax where $a \in A$ and $x \in X$. This is a submodule of M , and it is easy to verify that

$$(A + B)X = AX + BX \quad \text{and} \quad A(BX) = (AB)X$$

hold for all left ideals A and B . Note that $RX = X$ if X is a submodule of M . Taking $M = R$ in the above discussion shows that multiplication of left ideals is associative: $A(BC) = (AB)C$ for any left ideals A, B , and C .

An ideal A is called **nilpotent** if $A^n = 0$ for some $n \geq 1$; equivalently if any product of n elements of A is zero. The next result, proved in 1942 by Richard Brauer, characterizes the non-nilpotent, simple left ideals.

Lemma 9. Brauer's Lemma. *Let K be a simple left ideal of a ring R . Then either $K^2 = 0$ or $K = Re$ for some nonzero idempotent $e^2 = e \in K$.*

Proof. Assume that $K^2 \neq 0$, so that $Ka \neq 0$ for some $0 \neq a \in K$. Since $Ka \subseteq K$ is a left ideal, it follows that $Ka = K$ because $_RK$ is simple. In particular $ea = a$ for some $0 \neq e \in K$. Hence $e^2a = ea$, so $e^2 - e \in B = \{b \in K \mid ba = 0\}$. Moreover, B is a left ideal and $B \neq K$ because $Ka \neq 0$, so $B = 0$, again by simplicity. This means that $e^2 = e$, and so $Re \subseteq K$ because $e \in K$. But $e \neq 0$ so $Re = K$ by a third appeal to the simplicity of $_RK$. This is what we wanted. \blacksquare

We now turn to another property of rings that plays a prominent role in the Wedderburn–Artin theorem. A ring R is called **semiprime** if it satisfies the following equivalent conditions (the routine verifications are left to the reader):

- (1) *If $A^2 = 0$, A a left (or right, or two-sided) ideal, then $A = 0$.*
- (2) *If $A^n = 0$, $n \geq 1$, A a left (or right, or two-sided) ideal, then $A = 0$.*
- (3) *If $aRa = 0$ where $a \in R$, then $a = 0$.*

Hence, every ring with no nonzero nilpotent elements is semiprime, and the converse is true for commutative rings. A product $R = R_1 \times R_2 \times \cdots \times R_n$ of rings is semiprime if and only if each R_i is semiprime (using condition (3)). A matrix ring $M_n(R)$ is semiprime if and only if R is semiprime because the ideals of $M_n(R)$ all have the form $M_n(A)$ for some ideal A of R by Lemma 3 §3.3.

The next theorem gives several useful characterizations of when a ring R is semisimple as a left module over itself.

Theorem 4. *If R is a ring the following conditions are equivalent:*

- (1) *$_RR$ is semisimple*
- (2) *Every left R -module is semisimple.*
- (3) *Every left R -module is projective.*
- (4) *R is left artinian and semiprime.*

Proof. (1) \Rightarrow (2). Each module $_RM$ is an image of a free module, and free modules are semisimple by (1). So (2) follows from Corollary 1 of Theorem 1.

(2) \Rightarrow (3). Given a module M , let $\alpha : N \rightarrow M$ be an onto R -linear map. Then $\ker \alpha$ is a summand of N because N is semisimple by (2). This proves (3).

(3) \Rightarrow (1). Let L be any left ideal of R , and consider the coset map $\varphi : R \rightarrow R/L$. Then $L = \ker \varphi$ so L is a direct summand of $_RR$ because R/L is projective by (3). Hence $_RR$ is complemented, and so is semisimple by Theorem 1.

(1) \Rightarrow (4). First, R is left artinian by Lemma 4 because ${}_R R$ is finitely generated. Suppose A is an ideal of R with $A^2 = 0$. By Theorem 1 $R = A \oplus L$, L a left ideal, so $AL \subseteq A \cap L = 0$. Hence, $A = AR = A^2 + AL = 0 + 0 = 0$, so R is semiprime.

(4) \Rightarrow (1). If ${}_R R$ is not semisimple, let L be minimal among nonzero left ideals of R that are not semisimple (by (4)). Since $L \neq 0$, let $K \subseteq L$ be a simple left ideal, again by (4). Then $K^2 \neq 0$ because R is semiprime, so $K = Re$ for some $0 \neq e^2 = e$ by Brauer's lemma. Since $R = K \oplus R(1 - e)$ and $K \subseteq L$, we obtain

$$L = K \oplus (L \cap R(1 - e)) \text{ by the modular law (Theorem 2 §7.1).}$$

Hence, there are two cases: either $L \cap R(1 - e) = 0$ (in which case $L = K$ is simple) or $L \cap R(1 - e)$ is semisimple (by the minimality of L). Either way, L is semisimple, a contradiction. So ${}_R R$ is semisimple. \blacksquare

Note that we cannot replace “left artinian” by “left noetherian” in (4) of Theorem 4. In fact the ring \mathbb{Z} of integers is noetherian and semiprime, but it is not semisimple.

The first, and possibly the most important application of the theory of semisimple modules is to prove the following fundamental theorem: If R is a ring then ${}_R R$ is semisimple if and only if R is a finite direct product of matrix rings over division rings. This result was first proved in 1908 by Wedderburn for finite dimensional algebras. Then in 1927, Artin replaced the finite dimensional hypothesis by the descending chain condition on left ideals.

Theorem 5. Wedderburn–Artin Theorem. *The following conditions are equivalent for a ring R :*

- (1) ${}_R R$ is semisimple.
- (2) R_R is semisimple.
- (3) $R \cong M_{n_1}(D_1) \times M_{n_2}(D_2) \times \cdots \times M_{n_k}(D_k)$ for division rings D_i .

Moreover, the integers k , n_1, \dots, n_k in (3) are uniquely determined by R as are the division rings D_i up to isomorphism.

Proof. We need prove only (1) \Leftrightarrow (3) by the right–left symmetry of (3).

(1) \Rightarrow (3). Let H_1, H_2, \dots, H_m be the homogeneous components of ${}_R R$. Hence

$$R = H_1 \oplus H_2 \oplus \cdots \oplus H_m$$

by Theorem 2. Moreover, each H_i is an ideal of R (being fully invariant), so $R \cong H_1 \times H_2 \times \cdots \times H_m$ as rings by Theorem 7 §3.4 (and induction). So, by Wedderburn's theorem, it remains to show that H_i is left artinian and simple for each i . Write $H = H_i$ for convenience. The ring R is left artinian (by Lemma 4 since it is semisimple), so the same is true of $_R H$. But the left ideals of the ring H are exactly the left ideals of R that happen to be contained in H (verify). Hence ${}_H H$ is artinian. Finally, if $A \neq 0$ is an ideal of the ring H , then A is an ideal of R because $H = Re$ where $e^2 = e$ is central in R (verify). Hence, A is fully invariant in ${}_R R$. But then Theorem 2 shows that $A = \Sigma\{H_i \mid A \cap H_i \neq 0\}$, and it follows that $A = H$. Thus, the ring H is left artinian and simple, and (3) follows by Wedderburn's theorem.

(3) \Rightarrow (1). If D is a division ring, the ring $M_n(D) = K_1 \oplus K_2 \oplus \cdots \oplus K_n$ where K_j is the left ideal of all matrices with only column j nonzero. It is a routine verification (see Example 1) that each K_j is a simple left ideal of $M_n(D)$, so $M_n(D)$ is semisimple. Hence, (1) follows from (3).

Uniqueness. Suppose $R \cong M_{m_1}(B_1) \times M_{m_2}(B_2) \times \cdots \times M_{m_l}(B_l)$ is another such decomposition where each B_i is a division ring. Lemma 10 below shows that $k = l$ and, after relabeling, that $M_{n_i}(D_i) \cong M_{m_i}(B_i)$ for each $i = 1, 2, \dots, k$. We can now apply the uniqueness in Wedderburn's theorem (Theorem 3 §11.1). ■

The rings in Theorem 5 are called **semisimple** rings. The next result shows that this property is inherited by several related rings—the proofs are Exercises 10, 11, and 12.

Corollary. *If R is a semisimple ring, so also is every matrix ring $M_n(R)$, every factor ring R/A , and every corner ring eRe where $e^2 = e$.*

Note that subrings of semisimple rings need not be semisimple (consider $\mathbb{Z} \subseteq \mathbb{Q}$).

Lemma 10. *Let $R_1 \times R_2 \times \cdots \times R_k \cong S_1 \times S_2 \times \cdots \times S_l$ as rings, where each R_i and each S_j is a simple ring with a left composition series. Then $k = l$ and, after relabeling, $R_i \cong S_i$ for each i .*

Proof. Write $R = \prod_{i=1}^k R_i$ and $S = \prod_{j=1}^l S_j$. We may assume that $k \leq l$. Let $\sigma : R \rightarrow S$ be a ring isomorphism. The ideals of S are all of the form ΠA_j , where A_j is an ideal of S_j for each j (Exercise 4). Hence, since $\sigma(R_1)$ is a simple ideal of ΠS_j , it must be one of the S_j ; by relabeling assume that $\sigma(R_1) = S_1$. Similarly $\sigma(R_2) = S_j$ for some j , and $j \neq 1$ because $\sigma(R_1) \neq \sigma(R_2)$, σ is one-to-one, and $R_1 \cap R_2 = 0$. So, after relabeling, let $\sigma(R_2) = S_2$. Continue to obtain $S = \sigma(R_1) \times \cdots \times \sigma(R_k) \times S_{k+1} \times \cdots \times S_l$. Hence, $S \cong R \times (S_{k+1} \times \cdots \times S_l)$ so, since S and R have the same left composition length, the Jordan–Hölder theorem (Theorem 1 §11.1) shows that $S_{k+1} \times \cdots \times S_l$ has length zero. Hence, $k = l$ and $S_i = \sigma(R_i) \cong R_i$ for each i . This completes the proof. ■

Wedderburn's 1908 version of Theorem 5 was a breakthrough.¹³⁵ To quote Emil Artin, “This extraordinary result has excited the fantasy of every algebraist and still does so to this day. Very great efforts have been directed toward a deeper understanding of its meaning.”¹³⁶ When the Wedderburn–Artin theorem appeared in 1927 it was a landmark in algebra. It influenced a generation of ring theorists and has inspired many generalizations.

Wedderburn's theorem asserts that a left artinian simple ring is a matrix ring over a division ring. In 1945, Nathan Jacobson, and independently Claude Chevalley, extended Wedderburn's theorem by dropping the artinian hypothesis and showing that a simple ring with a simple left ideal must be isomorphic to a “dense” subring of the ring of endomorphisms of a vector space over a division ring. This result is called the **density theorem**.

Also in 1945, Jacobson showed that the intersection J of all the maximal left ideals of R always equals the intersection of the maximal right ideals of R . He then proved that J is the largest ideal with the property that $1 + a$ is a unit for all $a \in J$, extending work of Sam Perlis done in 1942 for finite dimensional algebras over a field. The ideal J is called the **Jacobson radical** of the ring R . It is known that,

¹³⁵ Maclagan Wedderburn, J.H. On hypercomplex numbers, *Proceeding of the London Mathematical Society, Series 2*, 6 (1908), 77–118.

¹³⁶ Artin, E. The influence of J.H.M Wedderburn on the development of modern algebra, *Bulletin of the American Mathematical Society* 56 (1950), 65–72.

if R is left (or right) artinian, the factor ring R/J is semisimple, and idempotents can be lifted modulo J in the sense that if $a^2 - a \in J$ then there exists $e^2 = e$ such that $a - e \in J$. In 1960 this led Hyman Bass to carry the theory further. He called a ring R **semiperfect** if R/J is semisimple and idempotents can be lifted modulo J , and he showed that many properties of left artinian rings carry over to these semiperfect rings.¹³⁷

Finally, the Wedderburn–Artin theorem shows that a left artinian semiprime ring is semisimple, and another natural question is what happens if we replace the left artinian condition by the requirement that the ring be left noetherian. In 1960, Alfred Goldie proved a fundamental structure theorem for the semiprime left noetherian rings. There is a way of embedding certain rings into a ring of **left quotients**, a noncommutative version of the construction of the field of quotients of an integral domain. Goldie showed that every left noetherian, semiprime ring has a *semisimple* ring of left quotients.

Exercises 11.2

1. Describe the semisimple \mathbb{Z} -modules, and the homogeneous ones.
2. Let R be a ring and let $a \in R$.
 - (a) If R is a semisimple ring and L is a left ideal, show that $L = Re$ for some $e^2 = e$ (so L is principal). [Hint: R is complemented.]
 - (b) In general, if $Ra = Re$ with $e^2 = e$, show that $aR = fR$ for some $f^2 = f$. [Hint: Show that $ata = a$ where $e = ta$, and use $f = at$.]
3. Let M_R be a right R -module, and write $E = \text{end}(M_R)$.
 - (a) Show that M is a left E -module via $\alpha \cdot x = \alpha(x)$ for all $\alpha \in E$ and $x \in M$.
 - (b) Show that $_EM$ is simple if and only if the only fully invariant submodules of M_R are 0 and M .
4. If A is an ideal of a product $R = R_1 \times R_2 \times \cdots \times R_n$ of rings, show that A has the form $A = A_1 \times A_2 \times \cdots \times A_n$, where A_i is an ideal of R_i for each i .
5. If N_1, \dots, N_m are all maximal submodules of a module M , show that $M/(\cap N_i)$ is semisimple.
6. Let $_RM = H_1 \oplus H_2 \oplus \cdots \oplus H_n$ where each H_i is fully invariant in M . Show that $\text{end } M \cong \text{end}(H_1) \times \text{end}(H_2) \times \cdots \times \text{end}(H_n)$ as rings.
7. If M is a finitely generated, semisimple module, show that $\text{end } M$ is a semisimple ring. [Hint: Preceding exercise and Lemma 3 §7.1.]
8. Let K be a simple module. If M is any module define $H_M(K)$ to be the sum of all submodules of M isomorphic to K , where $H_M(K) = 0$ if M has no submodule isomorphic to K . If $\alpha : M \rightarrow N$ is an R -linear map, show that $\alpha[H_M(K)] \subseteq H_N(K)$.
9. Show that every domain with a simple left ideal is a division ring. [Hint: Brauer's lemma.]
10. If R is semisimple show that $M_n(R)$ is semisimple for all $n \geq 1$. [Hint: Theorem 4.]
11. If R is semisimple show that R/A is a semisimple ring for all ideals A of R .
12. If R is semisimple show that eRe is semisimple if $e^2 = e \in R$. [Hint: Theorem 4.]
13. If R is semiprime and $e^2 = e \in R$, show that the following are equivalent: (1) Re is a simple left ideal; (2) eRe is a division ring; (3) eR is a simple right ideal. [Hint: Lemma 4 §11.1.]

¹³⁷Yes, there *are* perfect rings, indeed left and right perfect rings. They were also introduced by Bass, but a discussion of these rings is beyond the scope of this book.

14. If $R = \begin{bmatrix} F & F \\ 0 & F \end{bmatrix}$, F a field, let $e = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$. Show that eRe is a field, but Re is not simple (so the converse of Brauer's lemma is false). Show that eR is simple.
15. If R is semisimple, show that R is right and left noetherian. [Hint: Lemma 4.]
16. Let R be a semiprime ring.
- If L and M are left ideals, show that $LM = 0$ if and only if $ML = 0$.
 - If A and B are ideals show that $AB = 0$ if and only if $A \cap B = 0$.
 - If A is an ideal, and $r \in R$, show that $rA = 0$ if and only if $Ar = 0$.
17. If $R = R_1 \times R_2 \times \cdots \times R_n$, where the R_i are rings. Show that R is semiprime if and only if each R_i is semiprime.
18. If R is semiprime show that eRe is also semiprime for any $e^2 = e \in R$.
19. Show that a ring R is semiprime if and only if $M_n(R)$ is semiprime for some (any) $n \geq 1$.
20. A ring R is called **prime** if $AB = 0$, A, B ideals, implies that $A = 0$ or $B = 0$.
- Show that the commutative prime rings are the integral domains.
 - Show that a ring R is a prime and left artinian if and only if $R \cong M_n(D)$ for some $n \geq 1$ and some division ring D .
21. If P_1, P_2, \dots, P_n , are projective modules, show that $P_1 \oplus P_2 \oplus \cdots \oplus P_n$ is projective.
22. If M is a left module, define the **socle** of M , denoted $\text{soc}(M)$, to be the sum of all the simple submodules of M . (Take $\text{soc}(M) = 0$ if M contains no simple submodule). Show that
- $\text{soc}(M)$ is fully invariant in M .
 - If $N \subseteq M$ is a submodule then $\text{soc}(N) = N \cap \text{soc}(M)$.
 - If $M = N_1 \oplus N_2$ then $\text{soc}(M) = \text{soc}(N_1) \oplus \text{soc}(N_2)$.

Appendices

APPENDIX A COMPLEX NUMBERS

The set \mathbb{R} of real numbers has deficiencies. For example, the equation $x^2 + 1 = 0$ has no real root; that is, no real number u exists such that $u^2 + 1 = 0$. This type of problem also exists for the set \mathbb{N} of natural numbers. It contains no solution of the equation $x + 1 = 0$, and the set \mathbb{Z} of integers was invented to solve such equations. But \mathbb{Z} is also inadequate (for example, $2x - 1 = 0$ has no root in \mathbb{Z}), and hence the set \mathbb{Q} of rational numbers was invented. Again, \mathbb{Q} contains no solution to $x^2 - 2 = 0$, so the set \mathbb{R} of real numbers was created. Similarly, the set \mathbb{C} of complex numbers was invented that contains a root of $x^2 + 1 = 0$. More precisely, there is a complex number i such that

$$i^2 = -1.$$

However, the process ends here. The complex numbers have the property that every nonconstant polynomial with complex coefficients has a (complex) root. In 1799, at the age of 22, Carl Friedrich Gauss first proved this result, which is known as the **Fundamental Theorem of Algebra**. We give a proof in Section 6.6.

In this appendix, we describe the set \mathbb{C} of complex numbers. The set of real numbers is usually identified with the set of all points on a straight line. Similarly, the complex numbers are identified with the points in the Euclidean plane by labeling the point with cartesian coordinates (a, b) as

$$(a, b) = a + bi.$$

Then the set \mathbb{C} of **complex numbers** is defined by

$$\mathbb{C} = \{a + bi \mid a \text{ and } b \text{ in } \mathbb{R}\}.$$

When this is done, the resulting Euclidean plane is called the **complex plane**.

Each real number a is identified with the point $a = a + 0i = (a, 0)$ on the x -axis in the usual way, and for this reason the x -axis is called the **real axis**. The points

$bi = 0 + bi = (0, b)$ on the y -axis are called **imaginary numbers**, and the y -axis is called the **imaginary axis**.¹³⁸ The diagram shows the complex plane and several complex numbers.

Identification of the complex number $a + bi = (a, b)$ with the ordered pair (a, b) immediately gives the following condition for equality:

Equality Principle. $a = bi = a' + b'i$ if and only if $a = a'$ and $b = b'$.

For a complex number $z = a + bi$, the real numbers a and b are called the **real part** of z and the **imaginary part** of z , respectively, and are denoted by $a = \operatorname{re} z$ and $b = \operatorname{im} z$. Hence, the equality principle becomes as follows: Two complex numbers are equal if and only if their real parts are equal and their imaginary parts are equal.

With the requirement that $i^2 = -1$, we define **addition** and **multiplication** of complex numbers as follows:

$$(a + bi) + (a' + b'i) = (a + a') + (b + b')i,$$

$$(a + bi)(a' + b'i) = (aa' - bb') + (ab' + ba')i.$$

These operations are analogous to those for linear polynomials $a + bx$, with one difference: $i^2 = -1$. These definitions imply that complex numbers satisfy all the arithmetic axioms enjoyed by real numbers. Hence, they may be manipulated in the obvious fashion, except that we replace i^2 by -1 whenever it occurs.

Example 1. If $z = 2 - 3i$ and $w = -1 + i$,

$$\begin{aligned} z + w &= (2 - 1) + (-3 + 1)i = 1 - 2i, \\ z - w &= (2 + 1) + (-3 - 1)i = 3 - 4i, \\ zw &= (-2 - 3i^2) + (2 + 3)i = 1 + 5i, \\ \frac{1}{3}z &= \frac{1}{3}(2 - 3i) = \frac{2}{3} - i, \\ z^2 &= (2^2 + 9i^2) + 2(-6)i = -5 - 12i. \end{aligned}$$

Example 2. Find all complex numbers z such that $z^2 = -i$.

Solution. Write $z = a + bi$, where a and b are to be determined. Then the condition $z^2 = -i$ becomes

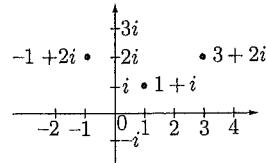
$$(a^2 - b^2) + 2abi = 0 + (-1)i.$$

Equating real and imaginary parts gives $a^2 = b^2$ and $2ab = -1$. The solution is

$$b = -a = \pm \frac{1}{\sqrt{2}}, \quad \text{so} \quad z = \pm \left(\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}i \right) = \pm \frac{1}{\sqrt{2}}(1 - i). \quad \square$$

Theorem 1 collects the basic properties of addition and multiplication of complex numbers. The verifications are straightforward and left to the reader.

¹³⁸As the terms *complex* and *imaginary* suggest, these numbers met with some resistance when they were first introduced. The names are misleading: These numbers are no more *complex* than the real numbers, and i is no more *imaginary* than -1 . Descartes introduced the term *imaginary numbers*.



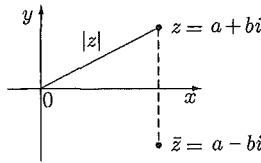
Theorem 1. If z , u , and w are complex numbers, then

- (1) $z + w = w + z$ and $zw = wz$.
- (2) $z + (u + w) = (z + u) + w$ and $z(uw) = (zu)w$.
- (3) $z + 0 = z$ and $z \cdot 1 = z$.
- (4) $z(u + w) = zu + zw$.

The following two notions are indispensable when working with complex numbers. If $z = a + bi$ is a complex number, the **conjugate** \bar{z} and the **absolute value** (or **modulus**) $|z|$ are defined by

$$\bar{z} = a - bi \quad \text{and} \quad |z| = \sqrt{a^2 + b^2}.$$

Thus, \bar{z} is a complex number and is the reflection of z in the real axis (see the diagram), whereas $|z|$ is a nonnegative real number and equals the distance between z and the origin. Note that the absolute value of a real number $a = a + 0i$ is $|a| = \sqrt{a^2 + 0^2} = \sqrt{a^2}$ using the definition of absolute value for complex numbers, which agrees with the absolute value of a regarded as a *real number*.



Theorem 2. Let z and w denote complex numbers. Then

- (1) $\bar{z} \pm \bar{w} = \bar{z} \pm \bar{w}$
- (2) $\bar{z}\bar{w} = \bar{z}\bar{w}$
- (3) $\overline{(\bar{z})} = z$
- (4) z is real if and only if $\bar{z} = z$
- (5) $z\bar{z} = |z|^2$
- (6) $|z| \geq 0$ and $|z| = 0$ if and only if $z = 0$
- (7) $|zw| = |z||w|$

Proof. (1) We prove (2), (5), and (7) and leave the rest to the reader. If $z = a + bi$ and $w = c + di$, we compute

$$\bar{z}\bar{w} = (a - bi)(c - di) = (ac - bd) - (ad + bc)i,$$

$$\bar{z}w = \overline{(a + bi)(c + di)} = \overline{(ac - bd) + (ad + bc)i} = (ac - bd) - (ad + bc)i,$$

which proves (2). Next, (5) follows from

$$z\bar{z} = (a + bi)(a - bi) = (a^2 + b^2) + (-ab + ba)i = a^2 + b^2 = |z|^2.$$

Finally (2) and (5) give

$$|zw|^2 = (zw)(\bar{z}\bar{w}) = zw\bar{z}\bar{w} = z\bar{z}w\bar{w} = |z|^2|w|^2.$$

Then (7) follows when we take positive square roots. ■

Let z be a nonzero complex number. Then (6) of Theorem 2 shows that $|z| \neq 0$, and so $z \left(\frac{1}{|z|^2} \bar{z} \right) = 1$ by (5). As a result, we call the complex number $(1/|z|^2)\bar{z}$ the **inverse** of z and denote it $z^{-1} = 1/z$, which proves Theorem 3.

Theorem 3. If $z = a + bi$ is a nonzero complex number, then z has an inverse given by

$$z^{-1} = \frac{1}{|z|^2} \bar{z} = \left(\frac{a}{a^2 + b^2} \right) - \left(\frac{b}{a^2 + b^2} \right) i.$$

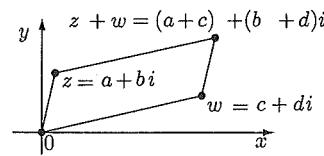
Hence, for real numbers, dividing by any nonzero complex number is possible. Example 3 shows how division is done in practice.

Example 3. Express $\frac{3+2i}{2+5i}$ in the form $a + bi$.

Solution. We multiply the numerator and denominator by the conjugate $2 - 5i$ of the denominator:

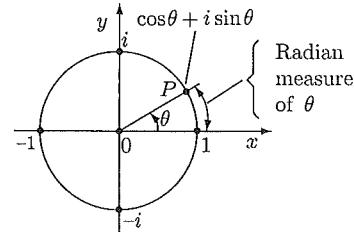
$$\frac{3+2i}{2+5i} = \frac{(3+2i)(2-5i)}{(2+5i)(2-5i)} = \frac{(6+10)+(4-15)i}{2^2+5^2} = \frac{16}{29} - \frac{11}{29}i. \quad \square$$

The addition of complex numbers has a geometric description. The diagram shows plots of the complex numbers $z = a + bi$ and $w = c + di$ and their sum $z + w = (a + c) + (b + d)i$. These points, together with the origin, form the vertices of a parallelogram, so we can find the sum $z + w$ geometrically by completing the parallelogram. This method is the **Parallelogram Law of Complex Addition** and is a special case of vector addition, as students of linear algebra will recognize.



The geometric description of complex multiplication requires that complex numbers be represented in polar coordinates. The circle with its center at the origin and radius 1 shown in the diagram below is called the **unit circle**. An angle θ measured counterclockwise from the real axis is said to be in **standard position**. The angle θ determines a unique point P on this circle. The **radian measure** of θ is defined to be the length of the arc from 1 to P . Hence, the radian measure of a right angle is $\pi/2$ radians and that of a full circle is 2π radians. We define the **cosine** and **sine** of θ (written $\cos \theta$ and $\sin \theta$) to be the x and y coordinates of P . Hence, P is the point $(\cos \theta, \sin \theta) = \cos \theta + i \sin \theta$ in the complex plane. These complex numbers $\cos \theta + i \sin \theta$ on the unit circle are denoted

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



A complete discussion of why we use this notation lies outside the scope of this book.¹³⁹

The fact that $e^{i\theta}$ is actually an *exponential* function of θ is confirmed by verifying that the law of exponents holds, that is,

$$e^{i\theta} e^{i\varphi} = e^{i(\theta+\varphi)} \quad \text{for any angles } \theta \text{ and } \varphi.$$

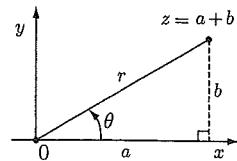
¹³⁹An entire theory exists for the study of functions such as e^z , $\sin z$, and $\cos z$, where z is a *complex* variable. Many theorems can be proved in this theory, including the Fundamental Theorem of Algebra mentioned previously.

This law is analogous to the exponent rule $e^a e^b = e^{a+b}$ for real exponents a and b , and it is an immediate consequence of the identities for $\sin(\theta + \varphi)$ and $\cos(\theta + \varphi)$:

$$\begin{aligned} e^{i\theta} e^{i\varphi} &= (\cos \theta + i \sin \theta)(\cos \varphi + i \sin \varphi) \\ &= (\cos \theta \cos \varphi - \sin \theta \sin \varphi) + i(\cos \theta \sin \varphi + \sin \theta \cos \varphi) \\ &= \cos(\theta + \varphi) + i \sin(\theta + \varphi) \\ &= e^{i(\theta+\varphi)}. \end{aligned}$$

We can now describe complex multiplication geometrically. We let $z = a + bi$ be any complex number. The distance r from z to 0 is the modulus $r = |z|$. If $z \neq 0$, it determines an angle θ , as shown in the diagram, called an **argument** of z . This angle is not unique ($\theta + 2\pi k$ would do as well for any $k = 0, \pm 1, \pm 2, \dots$) but, as the diagram clearly shows,

$$a = r \cos \theta \quad \text{and} \quad b = r \sin \theta$$



always hold. Hence, in any case

$$z = r(\cos \theta + i \sin \theta) = re^{i\theta}.$$

This expression is the **polar form** of the complex number z . The geometric description of complex multiplication follows from the law of exponents.

Theorem 4. Multiplication Rule. If $z = re^{i\theta}$ and $w = se^{i\varphi}$ are two complex numbers in polar form, then

$$zw = rse^{i(\theta+\varphi)}.$$

In other words, to multiply two complex numbers, simply multiply the absolute values and add the arguments. This method simplifies calculations and is valid for *any* arguments θ and φ .

Example 4. Multiply $(1 - i)(1 + \sqrt{3}i)$ by first converting the factors to polar form.

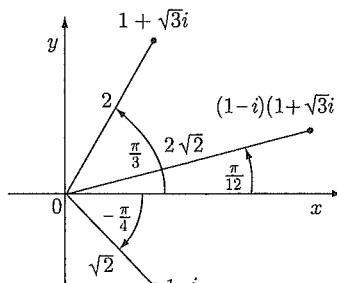
Solution. The polar forms (see the diagram) are

$$1 - i = \sqrt{2}e^{-\pi i/4}$$

and

$$1 + \sqrt{3}i = 2e^{\pi i/3}.$$

Hence, the multiplication rule gives



$$\begin{aligned} (1 - i)(1 + \sqrt{3}i) &= 2\sqrt{2}e^{i(\pi/3 - \pi/4)} \\ &= 2\sqrt{2}e^{i\pi/12} \\ &= 2\sqrt{2}(\cos \pi/12 + i \sin \pi/12). \end{aligned}$$

Of course, direct multiplication gives $(1-i)(1+\sqrt{3}i) = (\sqrt{3}+1) + (\sqrt{3}-1)i$, so equating real and imaginary parts gives the (somewhat unexpected) formulas

$$\cos\left(\frac{\pi}{12}\right) = \frac{\sqrt{3}+1}{2\sqrt{2}} \quad \text{and} \quad \sin\left(\frac{\pi}{12}\right) = \frac{\sqrt{3}-1}{2\sqrt{2}}. \quad \square$$

If $z = re^{i\theta}$ is given in polar form, $z^2 = r^2 e^{2i\theta}$ by the multiplication rule. Hence, $z^3 = (re^{i\theta})(r^2 e^{2i\theta}) = r^3 e^{3i\theta}$. In general, we have Theorem 5 for any $n \geq 1$ (we leave the proof for $n \leq 0$ as Exercise 15(b)). The name honors Abraham DeMoivre (1667-1754).

Theorem 5. DeMoivre's Theorem. If θ is any angle and $r > 0$, then $(re^{i\theta})^n = r^n e^{in\theta}$ for all integers n .

Example 5. Verify that $(-1 + \sqrt{3}i)^3 = 8$.

The polar form is $-1 + \sqrt{3}i = 2e^{2\pi i/3}$. Hence DeMoivre's theorem gives

$$(-1 + \sqrt{3}i)^3 = (2e^{2\pi i/3})^3 = 2^3 e^{2\pi i} = 2^3 \cdot 1 = 8. \quad \square$$

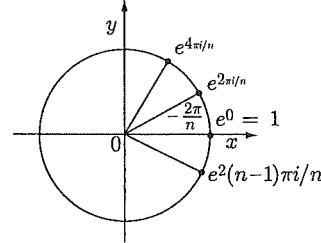
If $n \geq 1$, a complex number u is called an n th root of unity if $u^n = 1$. DeMoivre's theorem gives a way to find all possibilities (there are n). If we write $u = re^{i\theta}$ in polar form and use DeMoivre's theorem, the condition $u^n = 1$ becomes

$$r^n e^{in\theta} = 1e^{0i}.$$

Comparing absolute values gives $r^n = 1$, so $r = 1$ (because r is real and positive). However, the arguments may differ by integral multiples of 2π , so all we can conclude is that $n\theta = 2k\pi$, where k is an integer; that is,

$$\theta = \frac{2\pi k}{n}, \quad k \text{ an integer.}$$

These arguments give distinct values of u on the unit circle for $k = 0, 1, 2, \dots, n-1$, as shown in the diagram. But every choice of k yields a value of θ differing from one of these by a multiple of 2π , so they give all the possible roots. This proves Theorem 6.



Theorem 6. The n th roots of unity are $w_k = e^{2\pi ki/n}$ for $k = 0, 1, 2, \dots, n-1$.

We find these roots geometrically as the distinct points on the unit circle, starting at 1, that cut the circle into n equal sectors. Note that if $n = 2$, the roots are 1 and -1 , whereas the four 4th roots of unity are $1, i, -1$, and $-i$. In general, if we write $w = e^{2\pi i/n}$ then the n th roots of unity are $w^k = e^{2\pi ki/n}$ for each $k \geq 1$, so the n th roots of unity are just the powers of w :

$$1, w, w^2, \dots, w^{n-1}, \quad \text{where } w = e^{2\pi i/n}.$$

For this reason, $w = e^{2\pi i/n}$ is called a primitive n th root of unity. It follows easily (Exercise 16) that $1 + w + w^2 + \dots + w^{n-1} = 0$, that is, the sum of the n th roots of unity is zero.

Exercises A

1. Solve each equation for the real number x .
 - (a) $x - 4i = (2 - i)^2$
 - (b) $(2 + xi)(3 - 2i) = 12 + 5i$
 - (c) $(2 + xi)^2 = 4$
 - (d) $(2 + xi)(2 - xi) = 5$
2. Convert each expression to the form $a + bi$.

(a) $(2 - 3i) - 2(2 + 3i) + 9$ (c) $\frac{1+i}{2-3i} + \frac{1-i}{-2+3i}$ (e) i^{131} (g) $(1+i)^4$	(b) $(3-2i)(1+i) + 3+4i $ (d) $\frac{3-2i}{1-i} - \frac{3-7i}{2-3i}$ (f) $(2-i)^3$ (h) $(1-i)^2(2+i)^2$
--	---
3. In each case, find the complex number z .

(a) $iz - (1+i)^2 = 3 - i$ (c) $z^2 = -i$ (e) $2z + (1-7i) = (1+i)\bar{z}$	(b) $(i+z) - 3i(2-z) = iz + 1$ (d) $z^2 = 3 - 4i$
--	--
4. Let $\operatorname{re} z$ and $\operatorname{im} z$ denote the real and imaginary parts of z . Show that

(a) $\operatorname{im}(iz) = \operatorname{re} z$ (c) $z + \bar{z} = 2\operatorname{re} z$ (e) $\operatorname{re}(z+w) = \operatorname{re} z + \operatorname{re} w$, and $\operatorname{re}(tz) = t \cdot \operatorname{re} z$ if t is real (f) $\operatorname{im}(z+w) = \operatorname{im} z + \operatorname{im} w$, and $\operatorname{im}(tz) = t \cdot \operatorname{im} z$ if t is real	(b) $\operatorname{re}(iz) = -\operatorname{im} z$ (d) $z - \bar{z} = 2i\operatorname{im} z$
---	---
5. In each case, describe the graph of the equation, where z denotes a complex number

(a) $ z = 1$ (c) $z = i\bar{z}$ (e) $z = z $	(b) $ z - 1 = 2$ (d) $z = -\bar{z}$ (f) $\operatorname{im} z = m \cdot \operatorname{re} z$, m a real number
--	--
6. Verify $|zw| = |z| \cdot |w|$ directly for $z = a + bi$ and $w = c + di$.
7. Prove that $|w + z|^2 = |w|^2 + |z|^2 + w\bar{z} + \bar{w}z$ for all complex numbers w and z .
8. Show that $(1+i)^n + (1-i)^n$ is real for all integers $n \geq 1$.
9. (a) **Complex Distance Formula.** Show that $|z - w|$ is the distance between the complex numbers z and w .
 (b) **Triangle Inequality.** Show that $|z + w| \leq |z| + |w|$ for all complex numbers z and w . [Hint: Consider the triangle with vertices 0 , w , and $z + w$.]
10. Write each expression in polar form.

(a) $3 - 3i$ (d) $-4 + 4\sqrt{3}i$	(b) $-4i$ (e) $-7i$	(c) $-\sqrt{3} + i$ (f) $-6 + 6i$
---------------------------------------	------------------------	--------------------------------------
11. Write each expression in the form $a + bi$.

(a) $3e^{\pi i}$ (d) $\sqrt{2}e^{-\pi i/4}$	(b) $e^{7\pi i/3}$ (e) $e^{5\pi i/4}$	(c) $2e^{3\pi i/4}$ (f) $2\sqrt{3}e^{-2\pi i/6}$
--	--	---
12. Write each expression in the form $a + bi$.

(a) $(-1 + \sqrt{3}i)^2$ (c) $(1+i)^8$ (e) $(1-i)^6(\sqrt{3} + i)^3$	(b) $(1 + \sqrt{3}i)^{-4}$ (d) $(1-i)^{10}$ (f) $(\sqrt{3} - i)^9(2 - 2i)^5$
--	--
13. Use DeMoivre's theorem to show that

(a) $\cos 2\theta = \cos^2 \theta - \sin^2 \theta$; $\sin 2\theta = 2 \cos \theta \sin \theta$ (b) $\cos 3\theta = \cos^3 \theta - 3 \cos \theta \sin^2 \theta$; $\sin 3\theta = 3 \cos^2 \theta \sin \theta - \sin^3 \theta$	
--	--
14. Find all complex numbers such that

(a) $z^4 = -1$ (c) $z^3 = -27i$	(b) $z^4 = 2(\sqrt{3}i - 1)$ (d) $z^6 = -64$
------------------------------------	---
15. Let $z = re^{i\theta}$ in polar form.
 - (a) Show that $\bar{z} = re^{-i\theta}$ and $z^{-1} = \frac{1}{r}e^{-i\theta}$.
 - (b) Prove DeMoivre's theorem for $n \leq 0$.

APPENDIX B MATRIX ALGEBRA

Matrix algebra will be familiar to most readers as it is standard fare in beginning linear algebra courses. The new ingredient here is that the matrices we consider will have entries drawn from an arbitrary commutative ring R , rather than from the real numbers \mathbb{R} . A *ring* is an algebraic system in which there are two operations, addition and multiplication, for which the usual laws of arithmetic are valid (see Section 3.1). A ring R is called *commutative* if $ab = ba$ for all $a, b \in R$. Examples include the familiar number systems $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, and \mathbb{C} . It is worth noting that the set $M_2(\mathbb{R})$ of all 2×2 matrices over \mathbb{R} is a ring that is *not* commutative.

In this appendix, the standard results from linear algebra about matrices, adjugates, inverses, determinants, and so on, will be stated, with suitable minor modifications, over an arbitrary commutative ring R . We omit most proofs. When $R = \mathbb{R}$ many proofs from linear algebra remain valid, but this is often not the case. It is important to note that it is essential that R is commutative in many arguments, especially when dealing with inverses and determinants. Hence,

R denotes a commutative ring throughout Appendix B.

Matrix Algebra

A rectangular array of elements of R is called a **matrix** and the elements themselves are called the **entries** of the matrix. Thus,

$$A = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 5 & 6 \end{bmatrix} \quad C = \begin{bmatrix} 1 \\ -3 \\ 2 \end{bmatrix}$$

are matrices over \mathbb{Z} . The shape of a matrix depends on the number of **rows** and **columns**, and an $m \times n$ matrix is one with m rows and n columns. Two matrices are the same **size** if they have the same number of rows and the same number of columns. Thus, the preceding matrices A , B , and C are of size 2×2 , 2×3 , and 3×1 , respectively. An $n \times n$ matrix is called a **square matrix**.

The rows and columns of a matrix are numbered from the top down and from left to right, respectively. Then the entry in row i and column j of a matrix A is called the (i, j) -**entry** of A . If the (i, j) -entry is denoted a_{ij} , then A has the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

which usually is abbreviated as $A = [a_{ij}]$. Two $m \times n$ matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are **equal** (written $A = B$) if they have the same size and corresponding entries are equal, that is

$$[a_{ij}] = [b_{ij}], \quad \text{if and only if} \quad a_{ij} = b_{ij} \text{ for all } i \text{ and } j.$$

The set of all $m \times n$ matrices with entries from R is denoted $M_{mn}(R)$. For A and B in $M_{mn}(R)$, we obtain their **sum** $A + B$ by adding corresponding entries. If $A = [a_{ij}]$ and $B = [b_{ij}]$, this is

$$A + B = [a_{ij} + b_{ij}].$$

This addition enjoys many of the properties of numerical addition. For example, if A , B , and C are in $M_{mn}(R)$, then

$$A + B = B + A \quad \text{and} \quad A + (B + C) = (A + B) + C.$$

The matrix in $M_{mn}(R)$ each of whose entries is zero is called the **zero matrix** of size $m \times n$ and is denoted 0 (or 0_{mn} if the size must be emphasized). Clearly,

$$A + 0 = A, \quad \text{for all } A \text{ in } M_{mn}(R).$$

So 0 plays the role in $M_{mn}(R)$ that the number zero plays in \mathbb{Z} . We obtain the **negative** $-A$ of a matrix A in $M_{mn}(R)$ by negating every entry of A . Hence,

$$A + (-A) = 0, \quad \text{for all } A \text{ in } M_{mn}(R).$$

Finally we define **subtraction** by $A - B = A + (-B)$. If $A = [a_{ij}]$ and $B = [b_{ij}]$,

$$-A = [-a_{ij}] \quad \text{and} \quad A - B = [a_{ij} - b_{ij}].$$

With these definitions, the additive arithmetic in $M_{mn}(R)$ is entirely analogous to numerical arithmetic.

We also use the following notation: If A is a matrix and $r \in R$, the matrix rA is obtained by multiplying every entry of A by r . More formally

$$\text{If } A = [a_{ij}] \text{ then } rA = [ra_{ij}].$$

This is called **scalar multiplication** and one verifies that it enjoys the following useful properties for all scalars r, s and matrices A, B :

- (1) $r(A + B) = rA + rB$
- (2) $(r + s)A = rA + sA$
- (3) $r(sA) = (rs)A$
- (4) $1A = A$

Example 1. Given $A = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}$ in $M_{22}(\mathbb{Z})$, we have

$$2A - 5B = \begin{bmatrix} 2 & 4 \\ -2 & 0 \end{bmatrix} - \begin{bmatrix} 10 & 0 \\ 5 & 15 \end{bmatrix} = \begin{bmatrix} -8 & 4 \\ -7 & 15 \end{bmatrix}.$$

Example 2. If $A = \begin{bmatrix} 1 & -1 & 0 \\ 5 & 7 & -2 \end{bmatrix}$ and $B = \begin{bmatrix} 3 & 7 & -1 \\ 0 & 1 & 6 \end{bmatrix}$ in $M_{23}(R)$, find X in $M_{23}(R)$ such that $X + A = B$.

Solution. We proceed as in numerical arithmetic and subtract A from both sides:

$$X = B - A = \begin{bmatrix} 3 & 7 & -1 \\ 0 & 1 & 6 \end{bmatrix} - \begin{bmatrix} 1 & -1 & 0 \\ 5 & 7 & -2 \end{bmatrix} = \begin{bmatrix} 2 & 8 & -1 \\ -5 & -6 & 8 \end{bmatrix}. \quad \square$$

Multiplication of matrices is less natural than addition. To describe it, we define the **dot product** of a row matrix and a column matrix as follows:

$$[a_1 \ a_2 \ \cdots \ a_k] \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k.$$

Now let A be an $m \times k$ matrix and B be a $k \times n$ matrix, chosen so that the rows of A and the columns of B have the same number k of entries. Then the **product** AB is defined to be the $m \times n$ matrix whose

(i, j) -entry is the dot product of row i of A and column j of B .

Thus to compute the (i, j) -entry, go *across* the i th row of A and *down* the j th column of B and form the dot product. *Note:* If A is $m \times k$ and B is $k' \times n$, then

AB is defined only if $k = k'$, and then the product AB is $m \times n$.

Example 3. For $A = \begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 \\ 0 & 2 \\ -1 & 0 \end{bmatrix}$, compute AB and BA .

Solution. We write out the dot products explicitly.

$$AB = \begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 2 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 6+0-2 & 3-2+0 \\ 0+0-4 & 0+2+0 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ -4 & 2 \end{bmatrix},$$

$$BA = \begin{bmatrix} 2 & 1 \\ 0 & 2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 3 & -1 & 2 \\ 0 & 1 & 4 \end{bmatrix} = \begin{bmatrix} 6+0 & -2+1 & 4+4 \\ 0+0 & 0+2 & 0+8 \\ -3+0 & 1+0 & -2+0 \end{bmatrix} = \begin{bmatrix} 6 & -1 & 8 \\ 0 & 2 & 8 \\ -3 & 1 & -2 \end{bmatrix}. \quad \square$$

Example 4. If $A = \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}$, compute A^2 , AB , and BA .

Solution.
$$A^2 = \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

so $A^2 = 0$ can occur even when $A \neq 0$. Next

$$\begin{aligned} AB &= \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -3 & 12 \\ 2 & -8 \end{bmatrix}, \\ BA &= \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 6 & 9 \\ -4 & -6 \end{bmatrix} = \begin{bmatrix} -2 & -3 \\ -6 & -9 \end{bmatrix}. \end{aligned}$$

Hence $AB \neq BA$ is possible even though they are both the same size. \square

Example 4 shows that two familiar properties of numerical algebra fail for matrices. Hence, it is surprising to learn that the following property *does* hold.

Theorem 1. Let A , B , and C be of sizes $m \times p$, $p \times q$, and $q \times n$, respectively. Then $(AB)C = A(BC)$.

Proof. Write $A = [a_{ij}]$, $B = [b_{ij}]$, and $C = [c_{ij}]$. Then $AB = [x_{ij}]$ where we have $x_{ij} = \sum_{k=1}^p a_{ik}b_{kj}$, and $(AB)C = [y_{ij}]$ where $y_{ij} = \sum_{t=1}^q x_{it}c_{tj}$. Hence,

$$y_{ij} = \sum_{t=1}^q \left(\sum_{k=1}^p a_{ik}b_{kt} \right) c_{tj} = \sum_{t=1}^q \sum_{k=1}^p a_{ik}b_{kt}c_{tj} = \sum_{k=1}^p a_{ik} \left(\sum_{t=1}^q b_{kt}c_{tj} \right).$$

This last expression is the (i, j) -entry of $A(BC)$, and the theorem follows. Note that we needed the associativity of R to get $a_{ik}(b_{kt}c_{tj}) = a_{ik}b_{kt}c_{tj} = (a_{ik}b_{kt})c_{tj}$. ■

We express this result by saying that matrix multiplication is *associative* when the matrix sizes are such that the products involved are all defined.

The number 1 plays a neutral role in numerical multiplication in the sense that $1a = a$ and $a1 = a$ for every number a . The analogous role in matrix algebra is played by the **identity matrices** I_n . For each $n \geq 1$, the matrix I_n is defined to be the $n \times n$ matrix with 1s along the **main diagonal** (upper left to lower right), and 0s elsewhere. Thus,

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \dots$$

We use I without a subscript for the identity matrix when there is no need to emphasize the size. The reader can verify that the relations

$$AI = A \quad \text{and} \quad IB = B$$

hold whenever the matrix products are defined.

Note that $rA = (rI)A$ for all $r \in R$ and all matrices A . Moreover, since R is commutative we have (when the matrix multiplications are defined)

$$r(AB) = (rA)B = A(rB), \quad \text{for all } r \in R.$$

Square Matrices and Inverses

We are interested primarily in square matrices. For convenience, we use the notation

$$M_n(R) = M_{nn}(R), \quad \text{for any } n \geq 2.$$

If A and B lie in $M_n(R)$ then $A + B$ and AB are both in $M_n(R)$. Theorem 2 collects several properties of $M_n(R)$ for reference later.

Theorem 2. Let A, B , and C be matrices in $M_n(R)$. Then

- (1) $A + B = B + A$
- (2) $(A + B) + C = A + (B + C)$
- (3) $A + 0 = A$
- (4) $A + (-A) = 0$
- (5) $(AB)C = A(BC)$
- (6) $AI = A = IA$
- (7) $A(B + C) = AB + AC$ and $(B + C)A = BA + CA$

Proof. The only property not discussed previously is (7), and we leave the verification as Exercise 12. ■

The reader may have noted that Theorem 2 shows that $M_n(R)$ is a ring (noncommutative if $n \geq 2$). It is also noteworthy that Theorem 2 holds even if the ring R is not commutative.

If A is a square matrix, a matrix B is called an **inverse** of A if

$$BA = I \quad \text{and} \quad AB = I.$$

If it exists, this matrix B is uniquely determined by A . For if $AC = I$ also holds, then $AC = AB$ (both equal to I) so left multiplication by B gives $C = B$. A square matrix A is called **invertible** if it has an inverse, and in this case the (unique) inverse is denoted A^{-1} . Note that 0 has no inverse; in fact if $AB = 0$ for some $B \neq 0$ then A has no inverse.

Write R^n for the set of $n \times 1$ column matrices with entries from R . If $A \in M_n(\mathbb{R})$ it is known that A is invertible if and only if the system $AX = B$ of linear equations has a solution X for any $B \in \mathbb{R}^n$. This holds in general because $AB = I$ in $M_n(R)$ implies $BA = I$ (Exercise 13). On the other hand, if $A \in M_n(\mathbb{R})$ it is known that A is invertible if and only if the homogeneous linear system $AX = 0$ has only the trivial solution $X = 0$. But this condition fails for $R = \mathbb{Z}$ (consider $A = 2I$). Hence if A is a square matrix, we need other ways to determine when A^{-1} exists.

Again, it is well known that $A \in M_n(\mathbb{R})$ is invertible if and only if $\det A \neq 0$. If R is any commutative ring it turns out that the determinant of $A \in M_n(R)$ can be defined in such a way that A is invertible if and only if $\det A$ is a **unit** in R ($u \in R$ is a unit if $uv = 1 = vu$ for some $v \in R$). The idea is to define $\det A$ inductively. If $n = 1$ and $A = [a]$, this holds if we define

$$\det[a] = a.$$

If $n \geq 2$, assume inductively that $\det A$ has been defined for all $(n - 1) \times (n - 1)$ matrices over R . If A is $n \times n$ write A_{ij} for the $(n - 1) \times (n - 1)$ matrix obtained

from A by deleting row i and column j . Then, given $1 \leq i, j \leq n$ we define the (i, j) -cofactor of A by

$$c_{ij}(A) = (-1)^{i+j} \det A_{ij}.$$

With this, we **define** $\det A$ as follows:

$$\det A = a_{11} c_{11}(A) + a_{21} c_{21}(A) + \cdots + a_{n1} c_{n1}(A). \quad (*)$$

Then one shows that the following properties of determinants hold for rows:

- (a) If B is formed by multiplying a row of A by $u \in R$, then $\det B = u \det A$.
- (b) If A contains a row of zeros then $\det A = 0$.
- (c) If two rows of A are interchanged then $\det A$ changes sign.
- (d) If a multiple of a row of A is added to a different row, then $\det A$ is unchanged.
- (e) If two rows of A are identical then $\det A = 0$.

With this one can prove a result first given (for $R = \mathbb{R}$) in 1772 by Pierre Simon de Laplace.

Theorem 3. Cofactor Expansion Theorem. If $A = [a_{ij}]$ is an $n \times n$ matrix over a commutative ring R , then

- (1) $\det A = \sum_{i=1}^n a_{ij} c_{ij}(A)$, for each $j = 1, 2, \dots, n$,
- (2) $\det A = \sum_{j=1}^n a_{ij} c_{ij}(A)$, for each $i = 1, 2, \dots, n$.

Furthermore, (a)–(e) above hold for rows and for columns.

We omit the details.¹⁴⁰ The expressions in (1) and (2) of Theorem 3 are called the **cofactor expansions** along column j and row i , respectively. In words, to find the cofactor expansion of $\det A$ along a row or column, multiply the entries of the row (column) by the corresponding cofactors, and add the results.

The cofactor expansion theorem has many important consequences, and the following result is necessary to describe them. Given an $n \times n$ matrix $A = [a_{ij}]$, the **transpose** A^T of A is defined by

$$A^T = [b_{ij}], \quad \text{where } b_{ij} = a_{ji} \text{ for all } i \text{ and } j.$$

This is an $n \times n$ matrix obtained from A by interchanging elements symmetric about the main diagonal. For example, if $A = \begin{bmatrix} a & x \\ y & b \end{bmatrix}$ then $A^T = \begin{bmatrix} a & y \\ x & b \end{bmatrix}$. The familiar elementary properties of the transpose hold

- (1) $(A + B)^T = A^T + B^T$
- (2) $(rA)^T = rA^T$
- (3) $(AB)^T = B^T A^T$

when the matrix operations are defined (even for nonsquare matrices). We omit the proof of the following theorem.

¹⁴⁰The argument when $R = \mathbb{R}$ appears in Section 3.6 of Nicholson, W.K. *Linear Algebra with Applications*, 7th ed., McGraw-Hill Ryerson, 2012.

Theorem 4. Transpose Theorem. If A is a square matrix then $\det A^T = \det A$.

Surprisingly, the determinant function preserves matrix multiplication. Again we omit the proof.

Theorem 5. Multiplication Theorem. If A and B are $n \times n$ matrices then

$$\det(AB) = \det A \det B.$$

Given an $n \times n$ matrix $A = [a_{ij}]$ over a commutative ring R , we define the **adjugate** of A (also called the **classical adjoint** of A) as follows:

$$\text{adj } A = [c_{ij}(A)]^T.$$

This is also $n \times n$ and the cofactor expansion theorem (with some ingenuity) gives

Theorem 6. Adjugate Theorem. If A is $n \times n$ and $d = \det A$, then

$$A(\text{adj } A) = dI_n = (\text{adj } A)A.$$

Again we omit the details.

If it happens that $d = \det A$ is a unit in R , then (as R is commutative) the adjugate theorem gives

$$A[d^{-1}\text{adj}(A)] = I_n = [d^{-1}\text{adj}(A)]A,$$

from which A is invertible and $A^{-1} = d^{-1}\text{adj}(A)$. This proves part of

Theorem 7. Invertibility Theorem. If A is $n \times n$ then A is invertible if and only if $\det A$ is a unit in R . In this case $\det(A^{-1}) = (\det A)^{-1}$.

The rest of the proof follows from Theorem 5 (Exercise 14).

The case of 2×2 matrices is simple and so arises frequently in examples. The relevant facts are displayed next.

Example 5. If $n = 2$ and $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then we have $\text{adj } A = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ and $\det A = ad - bc$. Hence, if $ad - bc$ is a unit in R then $A^{-1} = (ad - bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ is a convenient formula for the inverse of A .

These results are enough to do most of the calculations in this book. A good reference source for this material (and much more) is the book by McDonald, B.R. *Linear Algebra over Commutative Rings*, Marcel Dekker Inc., New York, 1984.

Multilinear Approach

An $n \times n$ matrix A can be thought of as a row of columns in R^n , and it is instructive to view matrices that way. Hence, we write $A = [A_1, \dots, A_k, \dots, A_n]$, where A_k denotes column k of A . The determinant function $\det : M_n(R) \rightarrow R$ has two basic properties from this point of view.

First, \det is a **multilinear function** of the columns of a matrix, that is, it is a linear function of column k for each k (when we fix all other columns of A). More precisely, if we define

$\delta_k : R^n \rightarrow R$ by $\delta_k(X) = \det[A_1, \dots, X, \dots, A_n]$, for all $X \in R^n$, the requirement is that, for each k ,

$$\delta_k(rX + sY) = r\delta_k(X) + s\delta_k(Y) \text{ for all } r, s \in R \text{ and } X, Y \in R^n.$$

Second, \det is an **alternating** function of the columns of a matrix; that is if two distinct columns of A are interchanged to form B , then $\det B = -\det A$.

Before proceeding, write $X_n = \{1, 2, \dots, n\}$ and recall the *symmetric group* S_n (see Section 1.4) of all *permutations* of X_n , that is all bijections $\sigma : X_n \rightarrow X_n$. Each such σ is a product of *transpositions*, that is permutations that interchange two members of X_n . And σ is called *even* or *odd* according as it can be expressed as a product of an even, respectively odd, number of transpositions (the parity theorem ensures that this is well defined). The *sign* of σ is defined to be 1 or -1 according as σ is even or odd, and is written $(-1)^\sigma$.

Hence, the fact that \det is alternating means that if B is obtained from A by a series of column transpositions, then $\det B = (-1)^\sigma \det A$ where σ is the corresponding column permutation. With this it can be shown that $d = \det$ is the only multilinear, alternating function $d : M_n(R) \rightarrow R$ that satisfies $d(I) = 1$. Moreover, if we write $\sigma(k) = \sigma k$ when $\sigma \in S_n$ and $k \in X_n$, the following characterization of $\det A$ can be proved.

Theorem 8. If $A = [a_{ij}]$ is $n \times n$ then $\det A = \sum_{\sigma \in S_n} (-1)^\sigma a_{1\sigma 1} a_{2\sigma 2} \cdots a_{n\sigma n}$ where the sum ranges over all $n!$ elements σ of S_n .

Theorem 8 leads to other important properties of the determinant, and is often taken as the definition.

Exercises B

Throughout these exercises, R is assumed to be a commutative ring.

1. If A is square, $AB = 0$, and $B \neq 0$, show that A cannot be invertible.
2. (a) If $B = [B_1, \dots, B_k, \dots, B_n]$ where B_k is column k of B , show that $AB = [AB_1, \dots, AB_k, \dots, AB_n]$.
 (b) If $AX = 0$ for every $X \in R^n$, show that $A = 0$.
3. If $A = \begin{bmatrix} 0 & -5 & 1 \\ 3 & 0 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & -3 \\ 1 & -2 \\ 6 & -10 \end{bmatrix}$, show that $AB = I_2$ but $BA \neq I_3$.
4. Show that $A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$ is invertible in $M_2(R)$ if and only if a and c are both units.
 In this case find A^{-1} .
5. Find invertible 2×2 matrices A and B such that $A + B$ is not invertible.
6. Find a matrix X such that $AX = B$ if $A = \begin{bmatrix} 3 & -1 \\ 4 & 8 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & -1 & 2 \\ 0 & 3 & 6 \end{bmatrix}$.
7. If A and B are $n \times n$, show that $(A - B)(A + B) = A^2 - B^2$ if and only if $AB = BA$.
8. If $A^2 = 0$, show that $I + A$ is invertible and find $(I + A)^{-1}$ in terms of A .
9. If $A = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix}$, show that $A^3 = I$ and use this result to find A^{-1} in terms of A .
10. Show that $A = \begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$ satisfies $A^2 - 3A - 10I = 0$ and use this result to find A^{-1} in terms of A .
11. Let A and B denote $n \times n$ matrices.
 - (a) If A and B are invertible, show that A^{-1} and AB also are invertible, and find formulas for the inverses.
 - (b) If A , B and $A + B$ are invertible, show that $A^{-1} + B^{-1}$ is also invertible find a formula for the inverse.

- (c) If $I + BA$ is invertible show that $I + AB$ is invertible and find a formula for $(I + AB)^{-1}$. [Hint: $A(I + BA) = (I + AB)A$.]
12. Prove (7) of Theorem 2. (These expressions are called the **distributive laws**.)
13. (a) If $A, B \in M_n(R)$, show that $AB = I$ implies that $BA = I$.
- (b) Show that A is invertible if and only if $AX = B$ has a solution for every $B \in R^n$.
[Hint: Write $I = [E_1, \dots, E_k, \dots, E_n]$, use (a) Exercise 2(a).]
14. Prove Theorem 7 using Theorems 5 and 6.
15. Let E_{ij} denote the matrix in $M_n(R)$ with (i, j) -entry 1 and all other entries 0.
- (a) Show that $E_{ik}E_{mj} = \delta_{km}E_{ij}$ where $\delta_{km} = 0$ or 1 according as $k = m$ or $k \neq m$.
- (b) Show that $E_{11} + E_{22} + \dots + E_{nn} = I$.
- (c) Show that if $A = [a_{ij}] \in M_n(R)$ then $A = \sum_{i,j} a_{ij}E_{ij}$.
Here the E_{ij} are called **matrix units** in $M_n(R)$, and δ_{km} is called the **Kronecker delta**.

APPENDIX C ZORN'S LEMMA

The independent sets in a vector space are undeniably important, but the *largest* independent sets are the most important of all: they are the bases of the space. This theme that the “largest” objects of a given type are the most interesting is universal in mathematics, certainly in algebra. Zorn’s lemma is a set-theoretical principle that shows that maximal objects of various types exist and has become indispensable in many parts of mathematics. Clarifying what “maximal” means is best formulated using the concept of a partial ordering.

A **partial order** on a nonempty set P is a relation¹⁴¹ \leq on P that satisfies the following conditions (where x, y , and z denote elements of P) :

- P1 $x \leq x$ for all $x \in P$ (*reflexivity*).
- P2 If $x \leq y$ and $y \leq z$, then $x \leq z$ (*transitivity*).
- P3 If $x \leq y$ and $y \leq x$, then $x = y$ (*antisymmetry*).

A set P with a partial ordering is called a **partially ordered set** (**poset** for short). We say that (P, \leq) is a poset to assert that \leq is a partial ordering on the set P . Posets occur everywhere in mathematics; here are a few examples.

The following are easily verified to be partial orders:

Inclusion \subseteq on any nonempty collection of sets.

Divisibility $|$ on any nonempty set of positive integers.

The usual ordering \leq on the set \mathbb{R} of real numbers.

Nonempty subsets of a poset are again posets with the same partial order.

Other examples will occur later.

If \leq is a partial order on a set P , we write $x < y$ to mean $x \leq y$ and $x \neq y$. An element $m \in P$ is called **maximal** in P if there is no element x of P such that $m < x$, or equivalently,

If $m \leq x$, where $x \in P$, then $m = x$.

¹⁴¹See Section 0.4.

For example, if U is any nonempty set, let \mathcal{P} denote the set of proper subsets $X \subset U$. Then the maximal members of \mathcal{P} (under inclusion) are the sets $U \setminus \{a\}$ omitting one element $a \in U$. Here is a more algebraic example.

Example 1. If R is a ring, the maximal ideals of R (Section 3.3) are defined to be the maximal members of the poset $\mathcal{P} = \{A \neq R \mid A \text{ is an ideal of } R\}$ partially ordered by inclusion. If R is commutative, corollary 1 of Theorem 6 §3.3 shows that an ideal A is maximal if and only if the factor ring R/A is a field.

Zorn's lemma gives a condition that guarantees that maximal elements exist in certain posets. To state it, we need some terminology. Let (P, \leq) be a poset. If $X \subseteq P$ is a nonempty subset, an element $u \in P$ is called an **upper bound** for X if $x \leq u$ for every $x \in X$. Note that u need *not* be an element of X . For example, in the poset (\mathbb{R}, \leq) the interval $X = (0, 1) = \{r \in \mathbb{R} \mid 0 < r < 1\}$ has no maximal member, but any number $u \geq 1$ is an upper bound on X .

If we are looking for maximal elements in a poset P , choose $x_1 \in P$. If x_1 is maximal, we are done. Otherwise, $x_1 < x_2$ for some $x_2 \in P$. If x_2 is maximal, we are finished; otherwise $x_1 < x_2 < x_3$ for some $x_3 \in P$. Hence, either we find a maximal element at some stage, or we create elements x_1, x_2, x_3, \dots in P such that $x_1 < x_2 < x_3 < \dots$ is a strictly increasing sequence. So it is plausible that guaranteeing the existence of maximal elements in P will require some restriction on such ascending sequences from P . In 1935, Max Zorn found a condition on P that does this and is reasonably easy to verify in specific situations.

Two elements x and y in a poset P are called **comparable** if either $x \leq y$ or $y \leq x$, and the poset P is called a **chain** if any two elements are comparable. Thus, \mathbb{R} is a chain with respect to the usual partial ordering. A partially ordered set P is said to be **inductive** if every chain in P has an upper bound in P .

Theorem. Zorn's Lemma. Every inductive partially ordered set has a maximal element.

Note that to show that a poset is inductive, it is *not* enough to check that all *countable*¹⁴² chains $x_1 < x_2 < x_3 < \dots$ have upper bounds in P . For example, consider the set \mathcal{P} of all countable subsets of \mathbb{R} , partially ordered by inclusion. Then every countable chain in \mathcal{P} has an upper bound in \mathcal{P} , namely, its union (since unions of countable sets are again countable). But \mathcal{P} has no maximal member because such a maximal set would have to be \mathbb{R} itself, and \mathbb{R} is not a countable set.

Zorn's lemma has a wide variety of applications throughout mathematics; we give three examples from algebra. We begin with vector spaces. In Section 6.1, we proved that every finite dimensional vector space has a basis, but the proof that this holds in the infinite dimensional case requires Zorn's lemma. If V is a vector space over a field F , let X be a nonempty (possibly infinite) subset of V . Then X is called **independent** if every finite subset of X is independent (in the sense of Section 6.1), and X is said to **span** V if every element of V is a linear combination of (a finite number of) elements of X . Finally, X is called a **basis** of V if it is independent and spans V .

¹⁴²A set X is called **countable** if it can be enumerated by the set \mathbb{N} of natural numbers, that is, if $X = \{x_0, x_1, x_2, x_3, \dots\}$.

Example 2. If F is a field, show that every nonzero vector space V has a basis.

Solution. The idea is to show that any maximal independent set is a basis. Let \mathcal{I} denote the set of all independent sets in V and partially order \mathcal{I} by inclusion. We begin by using Zorn's lemma to show that \mathcal{I} has maximal members. First, \mathcal{I} is not empty since $\{v\}$ is in \mathcal{I} for all $v \neq 0$ in V . Suppose that $\mathcal{C} = \{X_i \mid i \in I\}$ is a chain in \mathcal{I} ; if $X = \bigcup_{i \in I} X_i$, we claim that X is independent. To see this, let $\{x_1, x_2, \dots, x_n\}$ be a finite subset of X . Then each x_k is in some X_i , so, since X_i form a chain, $\{x_1, x_2, \dots, x_n\} \subseteq X_m$ for some m . Hence, $\{x_1, x_2, \dots, x_n\}$ is independent (since X_m is in \mathcal{I}), as required. This shows that X is in \mathcal{I} and so is an upper bound for \mathcal{C} in \mathcal{I} .

Now Zorn's lemma shows that \mathcal{I} has a maximal member B , and we claim that B is a basis of V . Since it is independent (being in \mathcal{I}), it remains to prove that $\text{span}(B) = V$, where $\text{span}(B)$ consists of all linear combinations of vectors in B . Assume on the contrary that $v \notin \text{span}(B)$; we show that $\{v\} \cup B$ is independent, contradicting the maximality of B .

So let X be a finite subset of $\{v\} \cup B$; we must show that X is independent. If $X \subseteq B$, this follows because $B \in \mathcal{I}$. If $X \not\subseteq B$, then $v \in X$, so write $X = \{v, x_1, x_2, \dots, x_n\}$, $x_i \in B$. Let $a_0v + a_1x_1 + a_2x_2 + \dots + a_nx_n = 0$, $a_i \in F$. Then $a_0 = 0$ because otherwise $v \in \text{span}(B)$, contrary to our choice (since a_0^{-1} exists in the field F). Hence, $a_1x_1 + a_2x_2 + \dots + a_nx_n = 0$, so $a_i = 0$ for $i \geq 1$, as required. \square

It is worth noting that Example 2 is true for any division ring in place of F . In fact, most of the theory of vector spaces goes through for division rings.

The next two examples of Zorn's lemma come from ring theory. An additive subgroup L of a ring R is called a **left ideal** if $Ra \subseteq L$ for all $a \in L$, where $Ra = \{ra \mid r \in R\}$. Right ideals are defined similarly, and the ideals of R (see Section 3.3) are just the left and right ideals. A **maximal left ideal** M is defined to be a maximal member of $\{L \neq R \mid L \text{ is a left ideal}\}$, partially ordered by inclusion.

Example 3. If $L \neq R$ is any left ideal, then L is contained in a maximal left ideal.

Solution. Let $\mathcal{P} = \{X \mid X \text{ is a left ideal and } L \subseteq X \neq R\}$, partially ordered by inclusion. Then \mathcal{P} is not empty ($L \in \mathcal{P}$), so it suffices to show that \mathcal{P} contains a maximal element. By Zorn's lemma, it is enough to show that \mathcal{P} is inductive. Hence, let $\{X_i \mid i \in I\}$ be a chain from \mathcal{P} ; we show that $X = \bigcup_i X_i$ is an upper bound for $\{X_i\}$. Clearly, X is a left ideal containing L , and we claim that $X \neq R$. For if $X = R$, then $1 \in X$, say $1 \in X_k$ for some $k \in I$. Since X_k is a left ideal, it follows that $R = R1 \subseteq X_k \subseteq R$, so $X_k = R$, contrary to the fact that $X_k \in \mathcal{P}$. So X is an upper bound in \mathcal{P} on the chain $\{X_i\}$, as required. \square

As our last example of the use of Zorn's lemma, we prove a theorem that is of central importance in the theory of commutative rings. Let R be a commutative ring. An ideal P of R is called a **prime ideal** of R if R/P is an integral domain, or equivalently, if $rs \in P$, where $r, s \in R$, then either $r \in P$ or $s \in P$ (see Theorem 3 §3.3). Every commutative ring has at least one prime ideal by Example 3 (since maximal ideals are prime in any commutative ring).

An element $a \in R$ is said to be **nilpotent** if $a^n = 0$ for some $n \geq 1$, and the set $\text{nil}(R)$ of all nilpotents in a commutative ring R is called the **nil radical** of

R . It is easy to verify that $\text{nil}(R) \subseteq P$ for every prime ideal P , and hence that $\text{nil}(R) \subseteq \cap\{P \mid P \text{ is a prime ideal of } R\}$. The following example uses Zorn's lemma to show that this is in fact equality.

Example 4. If R is a commutative ring, $\text{nil}(R) = \cap\{P \mid P \text{ is a prime ideal of } R\}$.

Solution. Let $a \in P$ for every prime ideal P ; by the above remarks, we must show that a is nilpotent. So we assume that a is not nilpotent and show (using Zorn's Lemma) that $a \notin P$ for some prime ideal P . To this end, let

$$\mathcal{A} = \{A \mid A \text{ is an ideal of } R \text{ and } a^n \notin A \text{ for every } n \geq 1\}.$$

Then \mathcal{A} is not empty as $0 \in \mathcal{A}$ since a is not nilpotent. Suppose $\{A_i \mid i \in I\}$ is a chain from \mathcal{A} , and let $A = \bigcup_{i \in I} A_i$. Then A is an ideal and if $a^n \in A$, then $a^n \in A_k$ for some k , contradicting the fact that $A_k \in \mathcal{A}$. Hence, $a^n \notin A$ for each n , and so $A \in \mathcal{A}$. This shows that \mathcal{A} is inductive, so, by Zorn's lemma, let $P \in \mathcal{A}$ be a maximal member of \mathcal{A} . Then certainly $a = a^1 \notin P$, so it remains to show that P is a prime ideal. To that end, let $rs \in P$, where $r, s \in R$; we must show that $r \in P$ or $s \in P$. Suppose, on the contrary, that $r \notin P$ and $s \notin P$. Then

$$Rr + P = \{tr + p \mid t \in R \text{ and } p \in P\} \text{ is an ideal of } R,$$

and $P \subset Rr + P$ because $r \notin P$. Since P is maximal in \mathcal{A} , it follows that $Rr + P$ is not in \mathcal{A} , and hence that $a^n \in Rr + P$ for some $n \geq 1$, say $a^n = t_1r + p_1$, where $t_1 \in R$ and $p_1 \in P$. Similarly, since $s \notin P$, there exists $m \geq 1$ such that $a^m = t_2s + p_2$, where $t_2 \in R$ and $p_2 \in P$. But then

$$a^{m+n} = (t_1r + p_1)(t_2s + p_2) = (t_1t_2)rs + (t_1r)p_2 + (t_2s)p_1 + p_1p_2.$$

Hence, $a^{m+n} \in P$ because rs, p_1 , and p_2 are all in P . This contradiction completes the proof. \square

The proof of Zorn's lemma is difficult and will be omitted.¹⁴³ It requires (and is in fact equivalent to) the **Axiom of Choice**, which asserts that if \mathcal{S} is any family of nonempty sets, we can form a set containing one element of each of the sets in \mathcal{S} . The axiom was first proposed in 1904 by Ernst Zermelo. At first glance, it seems self-evident, but there may be infinitely many choices of elements to make. For example, if \mathcal{S} consists of all the bounded intervals from \mathbb{R} , we can choose (say) the midpoint of each interval; but if \mathcal{S} consists of all nonempty subsets of \mathbb{R} , then it is not clear how to make all the choices. Bertrand Russell illustrates the point as follows: If a man has infinitely many pairs of shoes, and infinitely many pairs of socks, then he can easily choose one shoe from each pair (choose the left one, say), but choosing one sock from each pair requires the axiom of choice.

Mathematician's attitudes about the axiom of choice vary from never using it to making no distinction between mathematics assuming the axiom and mathematics not assuming it. Irving Kaplansky takes a middle position: "I try to remember to make a note of it when I use it, but I do not hesitate to use it." Whatever your attitude, Kurt Gödel showed in 1940 that the axiom of choice is consistent with the other axioms of set theory; that is, it cannot be disproved using these axioms. Then in 1963 Paul Cohen showed that the axiom of choice is independent of the other axioms; that is, it cannot be proved from these axioms.

¹⁴³See Kaplansky, I. *Set Theory and Metric Spaces*, Boston: Allyn & Bacon, 1972, Section 3.3.

Exercises C

1. If M is a finitely generated module (see Section 6.1), show that every submodule $K \neq M$ is contained in a maximal submodule N (that is, a submodule $N \neq M$ such that if $N \subseteq X$, where $X \neq M$ is a submodule, then $N = X$).
2. Let $K \subseteq M$ be modules (see Section 6.1).
 - (a) Show that there exists a submodule N maximal such that $K \cap N = 0$.
 - (b) If N is as in (a), show that $(K + N) \cap X \neq 0$ for every submodule $X \neq 0$.
[Hint: Consider the cases $X \subseteq N$ and $X \not\subseteq N$ separately.]
3. Show that every commutative ring R contains a minimal prime ideal Q , that is a prime ideal Q such that, if $P \subseteq Q$ and P is a prime ideal, then $P = Q$.

APPENDIX D PROOF OF THE RECURSION THEOREM

If A is a set, a mapping $\alpha : \mathbb{N} \rightarrow A$ is called a **sequence** from A . Sequences are usually described as follows: If we write $\alpha(n) = a_n$ for each $n \in \mathbb{N}$, then the sequence is denoted $a_0, a_1, a_2, a_3, \dots, a_n, \dots$. The recursion theorem is concerned with **recursively defined sequences** wherein the first term a_0 is specified and the later terms are uniquely determined by the earlier ones. Such sequences are unique if they exist (see Theorem 3 §1.1); our task here is to prove existence.

Theorem. Recursion Theorem. Given a set A and $a \in A$, there is exactly one sequence $a_0, a_1, a_2, a_3, \dots, a_n, \dots$ from A that satisfies the following requirements:

- (1) $a_0 = a$.
- (2) For $n \geq 1$, the term a_n is uniquely determined by $a_0, a_1, a_2, \dots, a_{n-1}$.

Proof. For each $n \geq 1$, let $\beta_n : A^n \rightarrow A$ be a mapping. We want to show that a sequence a_0, a_1, a_2, \dots exists such that

$$a_0 = a \quad \text{and} \quad a_n = \beta_n(a_0, a_1, \dots, a_{n-1}), \quad \text{for each } n \geq 1.$$

The sequence is just a mapping $\alpha : \mathbb{N} \rightarrow A$ such that $\alpha(n) = a_n$, and we construct α as a set of ordered pairs in $\mathbb{N} \times A$ (see Section 0.3). Call a subset $\lambda \subseteq \mathbb{N} \times A$ “nice” if it satisfies the following two conditions:

- (a) $(0, a)$ is in λ .
- (b) If (k, x_k) is in λ for $k = 0, 1, \dots, n - 1$, so also is $(n, \beta_n(a_0, a_1, \dots, a_{n-1}))$.

$\mathbb{N} \times A$ is clearly “nice”. Let α denote the intersection of all “nice” subsets λ , that is

$$\alpha = \{(n, x) \mid (n, x) \in \lambda \text{ for every “nice” subset } \lambda\}.$$

It is a routine verification that α is itself “nice”.

Claim 1. Given $n \in \mathbb{N}$, there exists $x \in A$ such that (n, x) is in α .

Proof. If $n = 0$ then $(0, a)$ is in α because α is nice. If $n > 0$ let (k, x_k) be in α for $k = 0, 1, \dots, n - 1$. Write $x = \beta_n(x_0, \dots, x_{n-1})$. Then (n, x) is in every “nice” set λ by the definition of α , so (n, x) is in λ by (b). It follows that (n, x) is in α . This proves Claim 1.

Claim 2. If both (n, x) and (n, y) are in α then $x = y$.

Proof. If $n = 0$ suppose that $(0, y)$ is in α where $y \neq a$. Consider $\alpha' = \alpha \setminus \{(0, y)\}$. It suffices to show that α' is “nice” since this contradicts the choice of α . Clearly $(0, a)$ is in α' . If (k, x_k) is in α' for $k = 0, 1, \dots, n - 1$ then all these pairs are in α , so $(n, \beta_n(x_0, \dots, x_{n-1}))$ is also in α . But $n \neq 0$ so this is actually in α' . This shows that α' is “nice”. Now assume that Claim 2 is true for $k = 0, 1, \dots, n - 1$, and that (n, x) and (n, y) are both in α where $x \neq y$. Then $\alpha'' = \alpha \setminus \{(n, y)\}$ is “nice” just as before, contrary to the choice of α . This proves Claim 2.

Finally let $n \in \mathbb{N}$. Then (n, a_n) is in α for some $a_n \in A$ by Claim 1, and a_n is unique by Claim 2. Hence a_0, a_1, a_2, \dots is the desired sequence. \square

Bibliography

This list identifies some of the books that the interested reader can peruse for more information on the topics discussed in this book. The list is by no means complete.

GENERAL ABSTRACT ALGEBRA

- Birkoff, G. and MacLane, S. *A Survey of Modern Algebra*, 4th ed., New York: Macmillan, 1977.
- Cohn, P.M. *Algebra*, Vols. 1 and 2, New York: Wiley, 1974, 1977.
- Dummit, D.S. and Foote, R.M. *Abstract Algebra*, 3rd ed., New York: Wiley, 2004.
- Herstein, I.N. *Topics in Algebra*, 2nd ed., New York: Wiley, 1975.
- Hungerford, T.W. *Abstract Algebra*, 2nd ed., New York: Holt, Reinhart and Winston, 1974.
- Jacobson, N. *Basic Algebra*, Vols. 1 and 2, San Francisco: Freeman, 1974, 1980.
- Van der Waerden, B.L. *Algebra*, Vols. 1 and 2, 7th ed., New York: Ungar, 1970.

NUMBER THEORY

- Burton, D.M. *Elementary Number Theory*, Boston: Allyn & Bacon, 1980.
- Davenport, H. *Higher Arithmetic*, New York: Harper, 1960.
- Hardy, G.H. and Wright, E.M. *An Introduction to the Theory of Numbers*, 4th ed., Oxford: Clarendon Press, 1960.
- LeVeque, W.J. *Topics in Number Theory*, Vols. 1 and 2, Reading, MA: Addison-Wesley, 1956.
- Niven, I. and Zuckerman, H.S. *An Introduction to the Theory of Numbers*, New York: Wiley, 1980.

GROUP THEORY

- Hall, M. *The Theory of Groups*, New York: Macmillan, 1959.
- Kaplansky, I. *Infinite Abelian Groups*, 2nd ed., Ann Arbor: University of Michigan Press, 1969.
- Kargapolov, M.I. and Merzljakov, Ju.I. *Introduction to the Theory of Groups*, New York: Springer-Verlag, 1979.
- Kurosh, A.E. *The Theory of Groups*, New York: Chelsea, 1960.
- Ledermann, W. *Introduction to Group Theory*, Edinburgh: Oliver and Boyd, 1973.
- Macdonald, I.D. *The Theory of Groups*, London: Oxford University Press, 1968.
- Rose, J.S. *A Course on Group Theory*, Cambridge, England: Cambridge University Press, 1978.
- Rotman, J.J. *An Introduction to the Theory of Groups*, 3rd ed., Boston: Allyn & Bacon, 1984.

RING THEORY

- Atiyah, M.E. and MacDonald, I.G. *Introduction to Commutative Algebra*, Reading, MA: Addison-Wesley, 1969.
- Herstein, I.N. *Noncommutative Rings*, Carus Monograph 15, Washington, D.C.: Mathematical Association of America, 1968.
- Kaplansky, I. *Commutative Rings*, Chicago: University of Chicago Press, 1974.
- Lam, T.Y. *A First Course in Noncommutative Rings*, New York: Springer-Verlag, 1991.
- McCoy, N.H. *Rings and Ideals*, Carus Monograph 8, Washington, D.C.: Mathematical Association of America, 1948.
- McDonald, B.R. *Linear Algebra over Commutative Rings*, New York: Marcel Dekker, 1984.

FIELD THEORY

- Artin, E. *Galois Theory*, Notre Dame, Ind.: University of Notre Dame Press, 1944.
- Kaplansky, I. *Fields and Rings*, 2nd ed. (rev.), Chicago: University of Chicago Press, 1972.
- Niven, I. *Irrational Numbers*, Carus Monograph 11, Washington, D.C.: Mathematical Association of America, 1956.
- Rotman, J. *Galois Theory*, New York: Springer-Verlag, 1990.
- Stewart, I.N. *Galois Theory*, London: Chapman and Hall, 1973.

RELATED BOOKS

- Artin, E. *Geometric Algebra*, New York: Interscience, 1957.
- Curtis, C.W. and Reiner, I. *Representation Theory of Finite Groups and Associative Algebras*, New York: Wiley, 1962.
- Halmos, P.R. *Naive Set Theory*, New York: Springer-Verlag, 1974.
- Kaplansky, I. *Set Theory and Metric Spaces*, Boston: Allyn & Bacon, 1972.
- Lidl, R. and Pilz, G. *Applied Abstract Algebra*, New York: Springer-Verlag, 1984.

- MacWilliams, F.J. and Sloane, N.J.A. *The Theory of Error-Correcting Codes*, New York: Wiley, 1952.
- Solow, D. *How to Read and Do Proofs*, 2nd ed., New York: Wiley, 1990.
- Wilder, R.L. *Introduction to the Foundations of Mathematics*, New York: Wiley, 1952.

HISTORICAL

- Bell, E.T. *Men of Mathematics*, 2nd ed., New York: Simon and Schuster, 1962.
- Boyer, C.B. *A History of Mathematics*, New York: Wiley, 1968.
- Courant, R., and Robbins, R. *What is Mathematics*, Oxford: Oxford University Press, 1941.
- Kline, M. *Mathematical Thought from Ancient to Modern Times*, New York: Oxford University Press, 1972.
- Newman, J.R. *The World of Mathematics* (4 Vol.), New York: Simon and Schuster, 1956.
- Van der Waerden, B.L. *A History of Algebra*, New York: Springer-Verlag, 1985.

Selected Answers

EXERCISES 0.1 PROOFS

1. (a) If $n = 2k$, k an integer, then $n^2 = 4k^2$ is a multiple of 4. The converse is true: If $n^2 = 4k$, then n must be even because n odd implies n^2 odd.
(c) Verify that $2^3 - 6 \cdot 2^2 + 11 \cdot 2 - 6 = 0$ and that $3^3 - 6 \cdot 3^2 + 11 \cdot 3 - 6 = 0$. The converse is false: $1^3 - 6 \cdot 1^2 + 11 \cdot 1 - 6 = 0$ but 1 is not 2 or 3. Thus 1 is a counterexample.
2. (a) Either n is even or it is odd; that is, $n = 2k$ or $n = 2k + 1$. Then $n^2 = 4k^2$ or $n^2 = 4(k^2 + k) + 1$.
3. (a) If n is even, it cannot be prime unless $n = 2$ because, otherwise, 2 is a proper factor. The converse is false: 9 is an odd integer greater than 2, which is not prime.
(c) If $\sqrt{a} > \sqrt{b}$, then $(\sqrt{a})^2 > (\sqrt{b})^2$; that is $a > b$, contrary to hypothesis. The converse is true: If $\sqrt{a} \leq \sqrt{b}$, then $(\sqrt{a})^2 \leq (\sqrt{b})^2$; that is $a \leq b$.
4. (a) If $\sqrt{x+y} = \sqrt{x} + \sqrt{y}$, then $x+y = (\sqrt{x} + \sqrt{y})^2 = x + 2\sqrt{xy} + y$. Hence $\sqrt{xy} = 0$, from which $xy = 0$; therefore $x = 0$ or $y = 0$, contrary to hypothesis.
5. (a) $n = 11$ is a counterexample because then $n^2 + n + 11$ has 11 as a factor.

EXERCISES 0.2 SETS

1. (a) $\{x \mid x = 5k \text{ where } k \in \mathbb{Z}^+\}$
2. (a) $\{1, 3, 5, 7, \dots\} = \{2k+1 \mid k \in \mathbb{N}\}$ (c) $\{-1, 1, -3\}$ (e) $\{\} = \emptyset$
3. (a) Not equal: $-1 \in A$ but $-1 \notin B$ (c) Equal to $\{a, l, o, y\}$
(e) Not equal: $1 \in A$ but $1 \notin B$ (g) Equal to $\{-1, 0, 1\}$
4. (a) $\emptyset, \{2\}$ (e) $\{1\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$
5. (a) True. As $B \subseteq C$, each element of B (in particular, A) is an element of C .
(c) False. $A = \{1\}$, $B = C = \{\{1\}, 2\}$.
6. Every element of $A \cap B$ is in both A and B by definition, so $A \cap B \subseteq A$ and $A \cap B \subseteq B$. If $X \subseteq A$ and $X \subseteq B$, then $x \in X$ implies that $x \in A$ and $x \in B$; that is, $x \in A \cap B$. Hence, $X \subseteq A \cap B$.

11. (a) $(x, y) \in A \times (B \cap C)$ if and only if $x \in A$ and $y \in B \cap C$; if and only if $x \in A$ and $y \in B$, and $x \in A$ and $y \in C$; if and only if $(x, y) \in A \times B$ and $(x, y) \in A \times C$; if and only if $(x, y) \in (A \times B) \cap (A \times C)$. Hence $A \times (B \cap C)$ and $(A \times B) \cap (A \times C)$ have the same elements.

EXERCISES 0.3 MAPPINGS

1. (a) Not a mapping: $\alpha(1) = -1$ is not in \mathbb{N} .
 (c) Not a mapping: $\alpha(-1) = \sqrt{-1}$ is not in \mathbb{R} .
 (e) Not a mapping: $\alpha(6) = \alpha(2 \cdot 3) = (2, 3)$ and $\alpha(6) = \alpha(1 \cdot 6) = (1, 6)$.
 (g) Not a mapping: $\alpha(2)$ not defined.
2. (a) Bijective
 (c) Onto, but not one-to-one
 (e) One-to-one but not onto
 (g) One-to-one but not onto if $|B| \geq 2$
3. (a) If $c \in C$, then $c = \beta\alpha(a) = \beta[\alpha(a)]$ for some $a \in A$. As $\alpha(a) \in B$, β is onto.
 (c) If $\beta(b) = \beta(b_1)$, write $b = \alpha(a)$ and $b_1 = \alpha(a_1)$, where $a, a_1 \in A$. Then $\beta[\alpha(a)] = \beta[\alpha(a_1)]$; that is, $\beta\alpha(a) = \beta\alpha(a_1)$. Because $\beta\alpha$ is one-to-one, $a = a_1$, which yields $b = \alpha(a) = \alpha(a_1) = b_1$.
7. (a) $\alpha^{-1}(y) = \frac{1}{a}(y - b)$
 (c) $\alpha^{-1} = \alpha$
9. If $\beta\alpha = 1_A$, then α is one-to-one so, as $|A| = |B|$ is finite, α is also onto. Hence α^{-1} exists so $\alpha^{-1} = 1_A\alpha^{-1} = \beta\alpha\alpha^{-1} = \beta 1_B = \beta$. Then $\alpha\beta = \alpha\alpha^{-1} = 1_B$ and $\beta^{-1} = (\alpha^{-1})^{-1} = \alpha$.
11. $\varphi^{-1}(x, y) = \alpha_2$ where $\alpha_2(1) = x$ and $\alpha_2(2) = y$.
14. (b) \Rightarrow (c). If $\alpha\gamma = \alpha\delta$, then $\gamma = 1_A\gamma = (\beta\alpha)\gamma = \beta(\alpha\gamma) = \beta(\alpha\delta) = (\beta\alpha)\delta = 1_A\delta = \delta$.
15. (c) \Rightarrow (a) If $b_0 \in B - \alpha(A)$, choose $a_0 \in A$, and define $\beta : B \rightarrow B$ by:

$$\beta(b) = \begin{cases} b, & \text{if } b \neq b_0, \\ \alpha(a_0), & \text{if } b = b_0. \end{cases}$$
 Deduce $b_0 = \alpha(a_0)$ using (c).

EXERCISES 0.4 EQUIVALENCES

1. (a) Equivalence: $[1] = [0] = [-1] = \{1, 0, -1\}$, $[2] = \{2\}$, $[-2] = \{-2\}$.
 (c) Not an equivalence: $x \equiv x$ only if $x = 1$.
 (e) Not an equivalence: $1 \equiv 2$ but $2 \not\equiv 1$.
 (g) Not an equivalence: $x \equiv x$ is never true.
 (i) Equivalence: $[(a, b)] =$ the line with slope 3 through (a, b) .
2. (a) $A_{\equiv} = \{[(1, 1)], [(1, 2)], [(1, 3)], [(2, 3)], [(3, 3)]\}$
 (c) $A_{\equiv} = \{[(1, 1)], [(2, 1)], [(3, 1)]\}$
3. (a) Kernel equivalence of $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}$, where $\alpha(n) = n^2$; $\sigma[n] = |n|$.
 (c) Kernel equivalence of $\alpha : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, where $\alpha(x, y) = y$; $\sigma[(x, y)] = y$.
7. (a) Not well defined: $\alpha(2) = \alpha\left(\frac{2}{1}\right) = 2$ and $\alpha(2) = \alpha\left(\frac{4}{2}\right) = 4$.
 (c) Not well defined: $\alpha\left(\frac{1}{2}\right) = 3$ and $\alpha\left(\frac{1}{2}\right) = \alpha\left(\frac{2}{4}\right) = 6$.
10. (c) $|A_{\equiv}| = |Q| = n$ by (c) of the preceding exercise.

EXERCISES 1.1 INDUCTION

2. (e) $\frac{1}{\sqrt{1}} + \cdots + \frac{1}{\sqrt{k}} + \frac{1}{\sqrt{k+1}} \geq \sqrt{k} + \frac{1}{\sqrt{k+1}} = \frac{\sqrt{k^2+k+1}}{\sqrt{k+1}} \geq \frac{k+1}{\sqrt{k+1}} = \sqrt{k+1}$.
3. (c) If $3^{2k+1} + 2^{k+2} = 7m$, then $3^{2k+3} + 2^{k+3} = 7(9m - 2^{k+2})$.

7. Clear if $n = 1$. In general, such a $(k + 1)$ digit number must end in 4, 5, or 6, and those are 3^k of each by induction. We are done since $3 \cdot 3k = 3^{k+1}$.
10. (a) If $k \geq 2$ cents can be made up, there must be a 2-cent or a 3-cent stamp. In the first case, replace a 2-cent stamp by a 3-cent stamp; in the second case, replace a 3-cent stamp by two 2-cent stamps.
15. If p_1 is true and $p_R \Rightarrow p_{R+1}$, show $X = \{n \mid p_n \text{ is false}\}$ is empty.
17. If p_n is “ n has a prime factor”, then p_2 is true. Assume p_2, \dots, p_k are all true. If $k + 1$ is a prime, we are done. If $k + 1 = ab$ write $2 \leq a \leq k$ and $2 \leq b \leq k$, then a (and b) has a prime factor by strong induction. Thus, $k + 1$ has a prime factor.
18. (a) $a_n = 2(-1)^n$ (c) $a_n = \frac{1}{2}[1 + (-1)^n]$
 24. (a) Verify p_1 and p_2 . (c) Verify p_1, p_2, \dots, p_{10} .

EXERCISES 1.2 DIVISORS AND PRIME FACTORIZATION

1. (a) (a) $391 = 23 \cdot 17 + 0$ (c) $-116 = (-9)13 + 1$
 2. (a) $n/d = 134.293\dots$, so $q = 134$. Then $r = 113$.
 9. (a) $6 = 3 \cdot 72 - 5 \cdot 42$ (c) $3 = 1 \cdot 327 - 6 \cdot 54$
 (e) $29 = 0 \cdot 377 + 1 \cdot 29$ (g) $1 = -17 \cdot 72 - 7 \cdot (-175)$
 11. (a) If $d = xm + yn$, where $x, y \in \mathbb{Z}$, then $1 = x\frac{m}{d} + y\frac{n}{d}$.
 15. If $d = \gcd(m, n)$ and $d_1 = \gcd(m_1, n_1)$, then $d \mid m$ and $d \mid n$, so $d \mid m_1$ and $d \mid n_1$ by hypothesis. Thus $d \mid d_1$.
 19. If $d = \gcd(m, n)$ and $d_1 = \gcd(km, kn)$, then $d \mid m$ and $d \mid n$, so $kd \mid km$ and $kd \mid kn$. Hence $kd \mid d_1$. To show that $d_1 \mid kd$, write $km = qd_1$ and $kn = pd_1$. We have $d = xm + yn$ where x and $y \in \mathbb{Z}$, so $kd = xkm + ykn = xqd_1 + ypd_1$. Thus $d_1 \mid kd$.
 27. Let $d = \gcd(m, p^k)$. Then $d \mid p^k$ so $d = p^j$, $j \leq k$. Show that $j > 0$ contradicts $\gcd(m, p^k) = 1$.
 30. (a) $3^4 7^3$ (c) $11 \cdot 13 \cdot 17$ (e) 241
 31. (a) 5 and 16, 170 (c) 139 and 278
 33. (a) 25, 200 has 90 positive divisors.
 41. (a) $\gcd(28, 665, 22, 869) = 63$, and $\text{lcm}(28, 665, 22, 869) = 10,405,395$.

EXERCISES 1.3 INTEGERS MODULO n

1. (a) True (c) True (e) True (g) False
 2. (a) $k \equiv 2 \pmod{7}$ (c) $k \equiv 0 \pmod{9}$
 3. (a) 2, 5, 10 (c) 3
 8. (a) 7
 9. (a) 7
 15. One of $a, a + 1$ is even so $2 \mid a(a + 1)(a + 2)$; similarly, one of $a, a + 1, a + 2$ is a multiple of 3. Since $\gcd(2, 3) = 1$, it follows that $2 \cdot 3 = 6$ divides $a(a + 1)(a + 2)$.
 17. Compute \bar{a}^3 for $0 \leq a \leq b$.
 22. (a) $\overline{27}, x = \overline{33}$ (c) $\overline{11}, x = \overline{16}$
 25. (a) $x = \bar{8}, y = \bar{5}$ (c) No solution
 (e) $(x, y) = (\bar{0}, \bar{4}), (\bar{1}, \bar{6}), (\bar{2}, \bar{1}), (\bar{3}, \bar{3}), (\bar{4}, \bar{5}), (\bar{5}, \bar{0}), (\bar{6}, \bar{2})$.
 27. (a) $\bar{3}, \bar{6}$ (c) No solution

31. (1) \Rightarrow (2). Let $n = p^k a$ where p is a prime and $p \nmid a$. If $a > 1$ then $\gcd(n, a) = a > 1$, so \bar{a} has no inverse in \mathbb{Z}_n . By (1), let $\bar{a}^k = \bar{0}$ in \mathbb{Z}_n . Then $n \mid a^k$, so $p \mid a^k$, so $p \mid a$, a contradiction.

35. (a) Working modulo p , $x^2 = \bar{1}$ means $x^2 - \bar{1} = \bar{0}$. Thus $(x - \bar{1})(x + \bar{1}) = \bar{0}$ so $x = \bar{1}$ or $x = -\bar{1}$ by Theorem 7.

EXERCISES 1.4 PERMUTATIONS

1. (a) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 5 & 1 & 4 \end{pmatrix}$ (c) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 5 & 4 \end{pmatrix}$ (e) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 2 & 3 & 1 \end{pmatrix}$

3. (a) $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 3 & 1 \end{pmatrix}$ (c) $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{pmatrix}$ (e) $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{pmatrix}$

7. (a) 24

11. (a) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 2 & 6 & 1 & 9 & 4 & 5 & 7 & 3 \end{pmatrix}$

13. (a) $(1 \ 4 \ 8 \ 3 \ 9 \ 5 \ 2 \ 7 \ 6)$ (c) $(1 \ 2 \ 8)(3 \ 6 \ 7)(4 \ 9 \ 5)$
(e) $(1 \ 3 \ 8 \ 7 \ 2 \ 5)$

17. (a) $(1 \ 4 \ 3 \ 2)(5 \ 7 \ 6)$

18. Odd

19. (a) Even (c) Even (e) Odd

25. It suffices to show that any pair of transpositions is a product of 3-cycles. If k, l, m , and n are distinct, this follows from $(k \ l)^2 = \varepsilon$, $(k \ l)(m \ n) = (k \ m \ l)(k \ m \ n)$, and $(k \ l)(k \ m) = (k \ m \ l)$.

EXERCISES 2.1 BINARY OPERATIONS

1. (a) Not commutative or associative; no unity, so no units.

(c) Commutative, associative, unity is 0; if $a \neq 1$, $a^{-1} = \frac{a}{a-1}$.

(e) Not commutative, associative; no unity, so no units.

(g) Commutative, associative; no unity, so no units.

(i) Not commutative, associative, unity is $(1, 0, 1)$; if $x \neq 0 \neq z$, $(x, y, z)^{-1} = \left(\frac{1}{x}, \frac{-y}{xz}, \frac{1}{z}\right)$.

	a	b
a	a	b
b	b	a

7. $M \times N$ is commutative if and only if both M and N are commutative. (m, n) is a unit if and only if both m and n are units, and then $(m, n)^{-1} = (m^{-1}, n^{-1})$.

9. (a) $a^{24}a = a^{25} = (a^5)^5 = (b^5)^5 = b^{25} = b^{24}b = a^{24}b$. Cancel a 24 times

13. If $(uv)w = 1$, show that $(vw)u = 1_U$.

18. (2) \Rightarrow (1). If $(ab)^{-1} = x$ then $(ab)x = 1$, so $a(bx) = 1$. Then $(bx)a = 1$ by (2). So a is a unit. Similarly for b .

EXERCISES 2.2 GROUPS

1. (a) Only 0 has an inverse.

(c) Group; unity is -1 , a^{-1} is $-a - 2$.

(e) Not closed: $(1 \ 2)(1 \ 3) = (1 \ 3 \ 2)$ is not in G .

(g) Group; unity is 16; each element is self-inverse.

(i) $n \mapsto 2n$ has no inverse in G .

3. (a) First $ad = c, a^2 = d$ by the Corollary to

Theorem 6. Next $ba \neq b, a, d$; and

$$ba = c \Rightarrow b = ac = a(ba) = (ab)a = 1a = a,$$

a contradiction. So $ba = 1$. Then $bd = a$,

$$bc = d, b^2 = c. \text{ Next, } ca = b, cd = 1, c^2 = a,$$

$$cb = d. \text{ Finally, } da = c, db = a, dc = 1, d^2 = b.$$

8. (a) Every element σ satisfies $\sigma^2 = \varepsilon$.

13. α is onto because $\alpha(g^{-1}) = g$ for all $g \in G$; α is one-to-one because $g^{-1} = h^{-1}$ implies that $g = (g^{-1})^{-1} = (h^{-1})^{-1} = h$.

23. (a) If $g = g^{-1}$, then $g^2 = gg^{-1} = 1$; if $g^2 = 1$, then $g^{-1} = g^{-1}1 = g^{-1}g^2 = g$.

29. (a) We first establish left cancellation: If $gx = gy$ in G , then $x = y$. In fact, let $hg = e$. Then $gx = gy$ implies $x = ex = hgx = hgy = ey = y$. With this, the fact that $hg = e = e \cdot e = hge$ gives $g = ge$ by left cancellation. This shows that e is the unity. Finally, $h(gh) = (hg)h = eh = h = he$, so $gh = e$, again by left cancellation. Thus, h is the inverse of g .

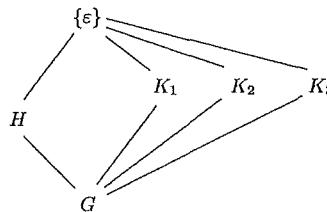
EXERCISES 2.3 SUBGROUPS

1. (a) No. $1 + 1$ is not in H . (c) No. $3^2 = 9$ is not in H .
 (e) No. $(1 \ 2)(3 \ 4) \cdot (1 \ 3)(2 \ 4) = (1 \ 4)(2 \ 3)$ is not in H .
 (g) Yes. $6 = 0$ is the unity. (i) Yes.
 6. (a) $1^2 = 1$. If $x = g^2, y = h^2$, then $xy = (gh)^2$ and $x^{-1} = (g^{-1})^2$.
 7. (a) $1 = g^0, g^k g^m = g^{k+m}$, and $(g^k)^{-1} = g^{-k}$; the subgroup test applies.
 8. (a) $1 = xx^{-1} \in \langle X \rangle$. Clearly $\langle X \rangle$ is closed. If $g = x_1^{k_1} \dots x_m^{k_m} \in X$, we obtain $g^{-1} = x_m^{-k_m} \dots x_1^{-k_1} \in X$. Hence, $\langle X \rangle$ is a subgroup; clearly $x = x^1 \in X$ so $X \subseteq \langle X \rangle$.

15. (a) $\{1\}$ and C_5 are the only subgroups of C_5 .



- (c) $\{\varepsilon\}, K_1 = \{\varepsilon, (1 \ 2)\}, K_2 = \{\varepsilon, (1 \ 3)\}, K_3 = \{\varepsilon, (2 \ 3)\}$,
 $H = \{\varepsilon, (1 \ 2 \ 3), (1 \ 3 \ 2)\}$, and S_3 .



17. \Rightarrow . If $H \not\subseteq K$, let $h \in H - K$. If $k \in K$, show $kh \notin K$, hence $kh \in H$, whence $k \in H$.

EXERCISES 2.4 CYCLIC GROUPS AND THE ORDER OF AN ELEMENT

1. (a) g, g^2, g^3, g^4 (c) $g, g^3, g^5, g^7, g^9, g^{11}, g^{13}, g^{15}$
 2. (a) 1, 2, 3, 4 (c) 1, 3, 5, 7, 9, 11, 13, 15
 4. (a) $\mathbb{Z}_7^* = \langle 3 \rangle$
 (c) \mathbb{Z}_{16}^* is not cyclic; $o(7) = o(9) = 2$; $o(3) = o(13) = o(5) = o(11) = 4$
 7. (a) 10 (c) 4

8. (a) $(1 \ 2 \ 3)(4 \ 5)$

9. (a)

(c)
(b) G $\langle g^2 \rangle$ $\langle g^4 \rangle$ $\{1\}$

(d)

 G $\langle g^3 \rangle$ $\langle g^9 \rangle$ $\langle g^6 \rangle$ $\langle 1 \rangle$

(e)

 G $\langle g^p \rangle$ $\{1\}$ $\langle g^q \rangle$ 11. (a) If $G = \langle a \rangle$ where $o(a) = n$, let $g = a^k$. Then $g^n = (a^k)^n = a^{kn} = (a^n)^k = 1^k = 1$.16. (a) $H = G$ (c) $H = \langle a^d \rangle$ (e) $H = \{(1, 1), (a, b), (a^2, b^2), (a^3, b^3), (a^3, b), (a, b^3), (a^2, 1), (1, b^2)\}$
 $= \{(a^k, b^m) \mid k + m \text{ is even}\}$ 17. (a) $X \subseteq Y \subseteq \langle Y \rangle$, $\langle Y \rangle$ a subgroup, so $\langle X \rangle \subseteq \langle Y \rangle$ by Theorem 8.25. Let $G = \langle g \rangle$ and $H = \langle h \rangle$ where $o(g) = m$, $o(h) = n$. As $|G \times H| = |G||H| = mn$, it suffices to show that $o((g, h)) = nm$. We have $(g, h)^{nm} = (g^{nm}, h^{nm}) = (1, 1)$. If $(g, h)^k = (1, 1)$, then $g^k = 1$ and $h^k = 1$, so $m \mid k$ and $n \mid k$. But $\gcd(n, m) = 1$, then implies $nm \mid k$ (Theorem 5 §1.2)), so $o(g, h) = mn$, as required.27. (a) If $A \subseteq B$, show that $g^a = g^{qb}$, $q \in \mathbb{Z}$. Since $|g| = \infty$, $a = qb$. Conversely, if $a = qb$, then $g^a \in B$, so $A \subseteq B$.

EXERCISES 2.5 HOMOMORPHISMS AND ISOMORPHISMS

8. (a) If $\alpha : \mathbb{Z} \rightarrow \mathbb{Z}$, let $\alpha(1) = m$. Then $\alpha(k) = \alpha(k \cdot 1) = k[\alpha(1)] = km$. Thus, α is multiplication by m , and each such map is a homomorphism $\mathbb{Z} \rightarrow \mathbb{Z}$.

12. (a) Yes. (c) No, not one-to-one. (e) Yes. (g) Yes. (i) Yes.

19. (a) If $z \in Z(G)$, then $\sigma(z) \in Z(G_1)$ because, given $g_1 = \sigma(g)$ in G_1 , $\sigma(z) \cdot g_1 = \sigma(zg) = \sigma(gz) = g_1 \cdot \sigma(z)$. Hence, $\sigma : Z(G) \rightarrow Z(G_1)$ is a mapping. It is one-to-one because σ is, and $\sigma(zw) = \sigma(z) \cdot \sigma(w)$ clearly holds. If $z_1 \in Z(G_1)$, let $z_1 = \sigma(z)$, $z \in G$. If $g \in G$, then $\sigma(gz) = \sigma(g) \cdot z_1 = z_1 \cdot \sigma(g) = \sigma(zg)$, so $gz = zg$ because σ is one-to-one. Thus, $z \in Z(G)$, and σ is onto.23. Write $w = e^{2\pi i/3} \in \mathbb{C}^\circ$. Suppose $\sigma : \mathbb{C}^\circ \rightarrow \mathbb{R}^*$ is an isomorphism, write $\sigma(w) = r$. Then $r^3 = [\sigma(w)]^3 = \sigma(w^3) = \sigma(1) = 1$, so $r = 1$, so $\sigma(w) = 1$, so $w = 1$, a contradiction.25. \mathbb{Z} is infinite cyclic, so $\mathbb{Q} \cong \mathbb{Z}$ is infinite cyclic too, say $\mathbb{Q} = \langle q \rangle = \{kq \mid k \in \mathbb{Z}\}$. In particular $q^2 = k_0q$, so $q = k_0 \in \mathbb{Z}$. Thus, $\mathbb{Q} = \{kk_0 \mid k \in \mathbb{Z}\} \subseteq \mathbb{Z}$, a contradiction.31. If $\sigma : G \rightarrow G$ is an automorphism, then $o(\sigma(a)) = 2$, so $\sigma(a) = a$. Because $\sigma(1) = 1$, $\sigma = 1_G$ and $\text{aut } G = \{1_G\}$.33. Let $G = \langle a \rangle$, $o(a) = \infty$. If $\sigma(a) = a^m$, $m \in \mathbb{Z}$, then $a = \sigma(a)^k = a^{mk}$. As $o(a) = \infty$ this gives $1 = mk$, whence $m = \pm 1$. If $m = 1$, then $\sigma = 1_G$; if $m = -1$ then $\sigma(g) = g^{-1}$ for all $g \in G$.

EXERCISES 2.6 COSETS AND LAGRANGE'S THEOREM

1. (a) $1H = \{1, a^4, a^8, a^{12}, a^{16}\}$
 $aH = \{a, a^5, a^9, a^{13}, a^{17}\}$
 $a^2H = \{a^2, a^6, a^{10}, a^{14}, a^{18}\}$
 $a^3H = \{a^3, a^7, a^{11}, a^{15}, a^{19}\}$

- $1K = \{1, a^2, a^4, a^6, a^8, a^{10}, a^{12}, a^{14}, a^{16}, a^{18}\}$
 $aK = \{a, a^3, a^5, a^7, a^9, a^{11}, a^{13}, a^{15}, a^{17}, a^{19}\}$
- (c) $0 + H = \{2k \mid k \in \mathbb{Z}\}$
 $0 + K = \{3k \mid k \in \mathbb{Z}\}$
 $1 + H = \{2k + 1 \mid k \in \mathbb{Z}\}$
 $1 + K = \{3k + 1 \mid k \in \mathbb{Z}\}$
 $2 + K = \{3k + 2 \mid k \in \mathbb{Z}\}$
5. (a) $a \equiv a$ because $a^{-1}a = 1 \in H$. If $a \equiv b$ then $b^{-1}a \in H$, and it follows that $a^{-1}b = (b^{-1}a)^{-1} \in H$, so $b \equiv a$. Finally, if $a \equiv b$ and $b \equiv c$ then $b^{-1}a \in H$ and $c^{-1}b \in H$, so $c^{-1}a = (c^{-1}b)(b^{-1}a) \in H$ and $a \equiv c$.
9. (a) The sets of positive and negative numbers.
(c) If $0 \leq t < 1$, $t + \mathbb{Z}$ is the set of numbers at distance t to the right of an integer.
10. (a) 6
12. (a) If $o(g) = m$, we show $m \mid 12$. We have $m \mid 12$ by Lagrange's theorem. If $m \neq 12$, then $m \mid 4$ or $m \mid 6$, so $g^4 = 1$ or $g^6 = 1$, contrary to hypothesis.
16. (a) If $1 = xm + yn$, where $x, y \in \mathbb{Z}$, then $g = g^1 = (g^m)^x(g^n)^y = 1^x 1^y = 1$.
25. (a) Because $a^kba^k = b$ implies that $a^{k+1}ba^{k+1} = aba = b$, it holds for $k \geq 0$. But $aba = b$ gives $b = a^{-1}ba^{-1}$, so $a^{-k}ba^{-k} = b$ follows for $k \geq 1$ in the same way.
31. If $|H : K| = n$, let Kh_1, \dots, Kh_n be the distinct cosets of K in H . This means $H = Kh_1 \cup \dots \cup Kh_n$, a disjoint union. Then $Hg \subseteq Kh_1g \cup \dots \cup Kh_ng$ is clear, and it is equality because $K \subseteq H$. Thus, each coset of H in G is the union of n K -cosets. If $|G : H| = m$ this gives $|G : K| = mn = |G : H| |H : K|$. Conversely, if $|G : K|$ is finite, then $|H : K|$ is clearly finite and $|G : H|$ is finite by the hint since each H -coset is a union of K -cosets.

EXERCISES 2.7 GROUPS OF MOTIONS AND SYMMETRIES

3. (a) If $\sigma = (1 \ 2 \ 3)$, the group of motions is $\langle \sigma \rangle = \{1, \sigma, \sigma^2\}$.
4. (a) If $\sigma = (1 \ 2 \ 3 \ 4)$, the group of motions is $\langle \sigma \rangle = \{1, \sigma, \sigma^2, \sigma^3\}$.
6. (a) If $\sigma = (1 \ 2 \ 3)(4 \ 5 \ 6)$ and $\tau = (1 \ 4)(2 \ 6)(3 \ 5)$, the group G of motions is $G = \{\varepsilon, \sigma, \sigma^2, \tau, \tau\sigma, \tau\sigma^2\} \cong D_3$.

EXERCISES 2.8 NORMAL SUBGROUPS

1. (a) Not normal (c) Normal
2. If $D_4 = \{1, a, a^2, a^3, b, ba, ba^2, ba^3\}$, where $|a| = 4$, $|b| = 2$, and $aba = b$, the normal subgroups are $\{1\}$, D_4 , $Z = \{1, a^2\} = Z(D_4)$, $H = \langle a \rangle$, $K_1 = \{1, a^2, b, a^2b\}$ and $K_2 = \{1, a^2, ba, ba^3\}$.
5. First aKa^{-1} is a subgroup, by Theorem 5 §2.3, and $aKa^{-1} \subseteq aHa^{-1} \subseteq H$ because $H \triangleleft G$. If $h \in H$, we must show $h(aKa^{-1})h^{-1} \subseteq aKa^{-1}$. We have $h^{-1}Kh = K$ as $K \triangleleft H$, so $h(aKa^{-1})h^{-1} = ha(h^{-1}Kh)a^{-1}h^{-1} = (hah^{-1})K(hah^{-1})^{-1} = K$ because $K \triangleleft G$.
11. Let H and K be subgroups of G with $|H| = p$ and $|K| = q$. Then $H \cap K = \{1\}$ by Lagrange's theorem. Moreover, $H \triangleleft G$ because it is unique of its order, and similarly $K \triangleleft G$. Hence, $G \cong H \times K$ by the Corollary to Theorem 6. Since p and q are primes, H and K are cyclic of relatively prime orders. Hence, $H \times K$ is cyclic by Exercise 25 §2.4.
15. (a) Conclude that $Ka = G \setminus K = Kb^{-1}$, so $ab \in K$.

17. (a) If $H = \langle a^d \rangle$, $d \mid n$, let $n = md$. Since ba^k is self-inverse for all k , we have $(ba^k)^{-1}a^{dt}(ba^k) = ba^k a^{dt}ba^k = b(a^{k+dt})ba^k = b \cdot b \cdot a^{-k-dt} \cdot a^k = b^2 a^{-k} a^{-dt} a^k = a^{-dt} \in H$.
24. (c) Let $G = C_2 \times C_2$, where $C_2 = \langle a \rangle$, $o(a) = 2$, and let $H = C \times \{1\}$. Then $H \triangleleft G$ because G is abelian. But $\sigma : G \rightarrow G$ given by $\sigma(x, y) = (y, x)$ is an automorphism and $\sigma(H) \not\subseteq H$. Thus, H is not characteristic in G .
26. (a) Write $K = \text{core } H = \cap_a aHa^{-1}$. Clearly $1 \in K$. If $g, g_1 \in K$, then $gg_1 \in aHa^{-1}$ for all a , so $gg_1 \in K$. Also, $g^{-1} \in a^{-1}H(a^{-1})^{-1}$ for all a , so $g \in K$. Hence, K is a subgroup. If $g \in G$, $k \in K$ then $gkg^{-1} \in g[(g^{-1}a)H(g^{-1}a)^{-1}]g^{-1} = aHa^{-1}$ for all a , as required.

EXERCISES 2.9 FACTOR GROUPS

1. (a) If $D_6 = \{1, a, \dots, a^5, b, ba, \dots, ba^5\}$, where $o(a) = 6$, $o(b) = 2$, and $aba = b$, then $K = \{1, a^3\}$ by Exercise 26 §2.6, and $D_6/K = \{K, Ka, Ka^2, Kb, Kba, Kba^2\}$.

D_6/K	K	Ka	Ka^2	Kb	Kba	Kba^2
K	K	Ka	Ka^2	Kb	Kba	Kba^2
Ka	Ka	Ka^2	K	Kba^2	Kb	Kba
Ka^2	Ka^2	K	Ka	Kba	Kba^2	Kb
Kb	Kb	Kba	Kba^2	K	Ka	Ka^2
Kba	Kba	Kba^2	Kb	Ka^2	K	Ka
Kba^2	Kba^2	Kb	Kba	Ka	Ka^2	K

- (c) $K(a, b) = K(1, b)$ because $(a, 1) \in K$. Thus, $G/K = \{K(1, b) \mid b \in B\}$. Moreover, $K(1, b) \cdot K(1, b_1) = K(1, bb_1)$, so the Cayley table is determined. *Remark:* The map $K(1, b) \mapsto b$ is an isomorphism $G/K \rightarrow B$.
3. (a) 6, 4, 3, 12
4. (a) 12
5. (a) 1, 2, 2, 2
7. If $0 < n < m$ in \mathbb{Z} , then $\mathbb{Z} + \frac{1}{n} \neq \mathbb{Z} + \frac{1}{m}$ because $\frac{1}{n} - \frac{1}{m} \notin \mathbb{Z}$. Hence, \mathbb{Q}/\mathbb{Z} contains the infinite set $\{\mathbb{Z} + \frac{1}{n} \mid n \geq 1\}$. Now let $\mathbb{Z} + \frac{m}{n}$ be any element of \mathbb{Q}/\mathbb{Z} . Then $n(\mathbb{Z} + \frac{m}{n}) = \mathbb{Z} + m = \mathbb{Z}$, so $\mathbb{Z} + \frac{m}{n}$ has finite order.
13. (a) If $z \in Z(G)$, then $Kz \in Z(G/K)$, so $z \in K$ by hypothesis. But then $z \in Z(K)$, so $z = 1$.
- (c) Given $z \in G$, let $(Kz)^{p^n} = K$. Then $z^{p^n} \in K$, so $(z^{p^n})^{p^m} = 1$; that is, $z^{p^{n+m}} = 1$. Hence $o(z)$ divides p^{n+m} , so $o(z) = p^k$ for some $k \geq 0$.
19. (a) $G' = \{1\}$
- (c) $D'_6 = \langle a^2 \rangle$ where $D_6 = \{1, a, \dots, a^5, b, ba, \dots, ba^5\}$, $o(a) = 6$, $o(b) = 2$, $aba = b$.
25. (a) $[Ka, Kb] = Ka \cdot Kb \cdot Ka^{-1} \cdot Kb^{-1}$
 $= K(aba^{-1}b^{-1})$
 $= K[a, b]$

EXERCISES 2.10 THE ISOMORPHISM THEOREM

4. (a) $1 \in \alpha^{-1}(X)$ because $\alpha(1) = 1 \in X$; if g and $h \in \alpha^{-1}(X)$, then $\alpha(gh) = \alpha(g)\alpha(h) \in X$ and $\alpha(g)^{-1} = \alpha(g^{-1}) \in X$, shows that $gh \in \alpha^{-1}(X)$ and $g^{-1} \in \alpha^{-1}(X)$. If $X \triangleleft \alpha(G)$, let $h \in \alpha^{-1}(X)$, $g \in G$. Then $\alpha(ghg^{-1}) = \alpha(g)\alpha(h)\alpha(g)^{-1} \in \alpha(g)X\alpha(g)^{-1} = X$, so $ghg^{-1} \in \alpha^{-1}(X)$. Hence, $\alpha^{-1}(X) \triangleleft G$.

8. (a) If $C_6 = \langle g \rangle$, $|g| = 6$, then the choice of $\alpha(g) \in K_4$ determines $\alpha : C_6 \rightarrow K_4$. If $\alpha(g) = 1$, then α is trivial. If $x \neq 1$ in K_4 , then $o(x) = 2$ and we define $\alpha_x : C_6 \rightarrow K_4$ by $\alpha_x(g^k) = x^k$. This mapping is well defined because

$$g^k = g^m \Rightarrow 6|(k-m) \Rightarrow 2|(k-m) \Rightarrow x^k = x^m.$$

Hence, α_x is a homomorphism and $\alpha_x(g) = x$. Thus, these are the only nontrivial homomorphisms.

(c) Let $D_3 = \{1, a, a^2, b, ba, ba^2\} = \langle a, b \rangle$, where

$o(a) = 3$, $o(b) = 2$, and $aba = b$, and let $C_4 = \langle c \rangle$, $o(c) = 4$. If $\alpha : D_3 \rightarrow C_4$ is a homomorphism, write $K = \ker \alpha$. Then $K \triangleleft D_3$, so $K = \{1\}$, $K = \langle a \rangle$ or $K = D_3$. Now $K = \{1\}$

is impossible as α is not one-to-one ($|D_3| = 6$ does not divide $|C_4| = 4$). If $K = D_3$, then α is trivial. So assume that α is not trivial. Then $\alpha(G) \cong G/K = \{K, bK\}$, so $\alpha(G)$ is the unique subgroup of C_4 of order 2: $\alpha(G) = \{1, c^2\}$. If $\varphi : G \rightarrow G/K$ is the coset map, there is an isomorphism $\sigma : G/K \rightarrow \alpha(G)$ such that $\alpha = \sigma\varphi$. Clearly, $\sigma(K) = 1$ and $\sigma(bK) = c^2$. Hence

$$\begin{aligned} \alpha(b^k a^m) &= \sigma\varphi(b^k a^m) = \sigma(b^k a^m K) = \sigma(b^k K) \\ &= \sigma[(bK)^k] = [\sigma(bK)]^k = c^{2k} \end{aligned}$$

This is the only nontrivial homomorphism.

10. (a) No. If $\alpha : S_3 \rightarrow K_4$ were onto, then $K_4 \cong S_3 / \ker \alpha$, and $|K_4| = 4$ would divide $|S_3| = 6$.
 (c) Yes. $S_3/A_3 \cong C_2$, say $\sigma : S_3/A_3 \rightarrow C_2$ is an isomorphism. If $\varphi : S_3 \rightarrow S_3/A_3$ is the (onto) coset map, then $\sigma\varphi : S_3 \rightarrow C_2$ is an onto homomorphism.
 13. Let G be simple. If $\alpha : G \rightarrow G_1$ is nontrivial, $\ker \alpha \neq G$, so $\ker \alpha = \{1\}$ by simplicity. So α is one-to-one and $G \cong \alpha(G) \subseteq G_1$. Conversely, if G_1 has a subgroup G_0 and $\sigma : G \rightarrow G_0$ is an isomorphism, then $\sigma : G \rightarrow G_1$ is a (one-to-one) homomorphism, which is nontrivial because $G_0 \neq \{1\}$ (it is simple).
 17. (a) If $g \in G'$, write $g = [a_1, b_1][a_2, b_2] \cdots [a_n, b_n]$, where $[a, b] = a^{-1}b^{-1}ab$. Then $\alpha(g) = \alpha[a_1, b_1] \cdots \alpha[a_n, b_n] = [\alpha(a_1), \alpha(b_1)] \cdots [\alpha(a_n), \alpha(b_n)] \in G'_1$.
 21. (a) Define $\alpha : \mathbb{C}^* \rightarrow \mathbb{R}^+$ by $\alpha(z) = |z|$ ($|z| > 0$ because $z \neq 0$). Then α is a homomorphism because $|zw| = |z||w|$, and $\ker \alpha = \{z \mid |z| = 1\} = \mathbb{C}^\circ$. Thus, the isomorphism theorem gives $\mathbb{C}^*/\mathbb{C}^\circ \cong \alpha(\mathbb{C}^*) \cong \mathbb{R}^+$.
 33. (a) \mathbb{Z}_4 has subgroups $H = \{0\}$, $\{0, 2\}$, and \mathbb{Z}_4 . Hence $\mathbb{Z}_4/H \cong \mathbb{Z}_4, \mathbb{Z}_2, \{1\}$, so these are the possible images.

$$\begin{array}{ccc} G & \xrightarrow{\alpha} & G_1 \\ \varphi \downarrow & \nearrow \sigma & \\ G/K & & \end{array}$$

EXERCISES 2.11 AN APPLICATION TO BINARY LINEAR CODES

1. (a) 5 (c) 6
2. (a) 3 (c) 7
7. (a) Detects 3, corrects 1.
11. (a) As $k = 3$ and $t = 2$, n must satisfy $\binom{n}{0} + \binom{n}{1} + \binom{n}{2} \leq 2^{n-3}$. If $n = 3, 4, \dots, 8$, this expression reads $7 \leq 1$, $11 \leq 2$, $16 \leq 4$, $22 \leq 8$, $29 \leq 16$, and $37 \leq 32$. Hence $n \geq 9$. [Note: For $n = 9$, it reads $46 \leq 64$.]
15. (a) If C is a $(4, 2)$ -code that corrects one error, the weight of C must be at least 3 so that the nonzero words in C are contained in $\{1111, 1110, 1101, 0111\}$. But the sum of any two of these words is not in the set.

20. (a) $G = [1 \ 1 \ 1 \ 1 \ 1]$

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(c) $G = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

21. (a) $\{0000, 1011, 0100, 1111\}$

(c) $\{000000, 100101, 010110, 001001, 110011, 101100, 011111, 111010\}$

EXERCISES 3.1 EXAMPLES AND BASIC PROPERTIES

1. (a) Not an additive group.

(c) $h(f+g) \neq hf+hg$ can happen (try $h(x) = x^2$).

3. (a) $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} = \begin{bmatrix} aa'+bc' & ab'+bd' \\ ca'+dc' & cb'+dd' \end{bmatrix} \in S$ because the column sums are $(aa'+bc') + (ca'+dc') = (a+c)a' + (b+d)c' = (a+c)(a'+c')$
 $(ab'+bd') + (cb'+dd') = (a+c)b' + (b+d)d' = (a+c)(b'+d') = (a+c)(a'+c')$.

The rest of the subring test is routine.

7. $\left\{ \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \mid a \in Z(R) \right\}$

14. Compute $(1+sr)[1-s(1+rs)^{-1}r]$.

15. (3) \Rightarrow (1). Let e be the unique right unity. Given $b \in R$, show that $r(e+eb-b)=r$ for all $r \in R$. Now use the uniqueness.

16. (a) $\mathbb{Z}1_R$ is a subring by Theorem 5. It is centered because $s \cdot (k1_n) = ks = (k1_n)_s$ for all $s \in R$ by Theorem 2.

18. (a) $\text{lcm}(m, n)$ (c) 0

21. (a) $(1-2e)^2 = 1 - 4e + 4e^2 = 1$

22. (a) If $a = (1-e)re$ then the fact that $e^2 = e$ gives $ea = 0$ and $ae = a$. It follows that $a^2 = (ae)a = a(ea) = a \cdot 0 = 0$.

23. (4) \Rightarrow (1). If $r \in R$, $a = (1-e)re$ is nilpotent so $u = 1+a$ is a unit and so commutes with e by (4). Conclude that $re = ere$.

29. (a) Units = $\{1, -1\}$; nilpotents = $\{0\}$; idempotents = $\{0, 1\}$

(c) Units: $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix};$

Nilpotents: $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix};$

Idempotents: $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$

36. (a) If $\sigma : \mathbb{C} \rightarrow \mathbb{R}$ is an isomorphism, then $a = \sigma(i)$ satisfies $a^2 = -1$, a contradiction.

(c) If $\mathbb{Z} \cong \mathbb{Q}$, then \mathbb{Z} is a division ring, a contradiction.

EXERCISES 3.2 INTEGRAL DOMAINS AND FIELDS

1. (a) 1, -4

(c) 0, 1

3. Idempotents = $\{0, 1\}$; nilpotents = $\{0\}$

7. If $ab = 0$ show that $(ba)^2 = 0$.

9. Try \mathbb{Z}_p for various primes p .

15. If $z \in Z(R)$ and $za = 1$, showing that $a \in Z(R)$ is sufficient. Given $r \in R$, $(ra - ar)z = raz - arz = r \cdot 1 - 1 \cdot r = 0$, so (as $za = 1$) $ra - ar = 0$.

16. (a) If $o \neq a \in K$ and $ab = 1$ with $b \in K$, conclude that $b = a^{-1}$ where a^{-1} is the inverse in F .

19. $\mathbb{Q}(\sqrt{2})$ is a subfield of R by Example 4. If F is any subfield of R then $\mathbb{Z} \subseteq F$ (because $1 \in R$), and hence $\mathbb{Q} \subseteq F$ (because $\frac{n}{m} = nm^{-1} \in F$ for all $n, m \neq 0$ in \mathbb{Z}). If also $\sqrt{2} \in F$, this means $r + s\sqrt{2} \in F$ for all $r, s \in \mathbb{Q}$. Thus $\mathbb{Q}(\sqrt{2}) \subseteq F$.

23. If $R = \{r_1, r_2, \dots, r_n\}$ and $0 \neq a \in R$, then ar_1, ar_2, \dots, ar_n are distinct ($ar_i = ar_j$ implies $r_i = r_j$ as $a \neq 0$). Hence $\{ar_1, ar_2, \dots, ar_n\}$ has n elements, and so equals R . Hence, $1 = ar_i$ for some i .

26. (a) If $\frac{r}{u} = \frac{r'}{u'}$ and $\frac{s}{v} = \frac{s'}{v'}$ show $(rs)(u'v') = (ru')(sv') = (r'u)(s'r) = (r's')(uv)$.

29. (a) If $r = i$ and $s = 1$ in \mathbb{C} , consider $a = r + sw$ in $\mathbb{C}(\omega)$. Then $aa^* = r^2 + s^2 = 0$, but $a \neq 0$ and $a^* \neq 0$ in $\mathbb{C}(\omega)$. Thus, $\mathbb{C}(\omega)$ is not a field. In $\mathbb{Z}_5(\omega)$ let $a = 1 + 2w$. Then $aa^* = 1^2 + 2^2 = 0$, and $a \neq 0 \neq a^*$. So $\mathbb{Z}_5(\omega)$ is not a field. However, $\mathbb{Z}_7(\omega)$ is a field. If $a = r + si \neq 0$ in $\mathbb{Z}_7(\omega)$ then $aa^* = r^2 + s^2$ and it suffices to show $r^2 + s^2 \neq 0$ in \mathbb{Z}_7 . Suppose $r^2 + s^2 = 0$. If $r = 0$ or $s = 0$ then $a = 0$, contrary to hypothesis. Thus $r \neq 0 \neq s$. Then $0 = s^{-1}(r^2 + s^2) = (s^{-1}r)^2 + 1$ so $(s^{-1}r)^2 = -1$ in \mathbb{Z}_7 . This is not the case because $0^2 = 0$, $1^2 = 1 = 6^2$, $2^2 = 4 = 5^2$, $3^2 = 2 = 4^2$ in \mathbb{Z}_7 .

32. (a) If q is a unit in $\mathbb{H}(R)$, then $1 = N(1) = N(qq^{-1}) = N(q)N(q^{-1})$, so $N(q)$ is a unit in R . Conversely, $qq^* = N(q)$ shows $q^{-1} = N(q)^{-1}q^*$ if $N(q) \in R^*$.

EXERCISES 3.3 IDEALS AND FACTOR RINGS

23. (a) 0 (c) $2R = \{0, 2, 4, 6, 8\}$
25. (c) If u is a unit then $u \in Ru$ implies $Ru = R$ by Theorem 2. Conversely, if $Ru = R$ then $1 \in Ru$, say $1 = vu$, $v \in R$. Hence, u is a unit (R is commutative).
33. (c) In $M_2(\mathbb{R})$, $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is nilpotent. If $B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ then $BA = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ is not nilpotent.
34. (c) Write $\mathbb{Z}_{p^n} = \mathbb{Z}/p^n\mathbb{Z}$. If $A = \mathbb{Z}/k\mathbb{Z} \neq 0$ is an ideal of \mathbb{Z}_{p^n} , then $p^n\mathbb{Z} \subseteq k\mathbb{Z}$, so $k \mid p^n$. Hence, $k = p^t$ for $t \leq n$, so $A \subseteq M$ where $M = \mathbb{Z}/p\mathbb{Z}$. It follows that M is the unique maximal ideal of \mathbb{Z}_{p^n} , so \mathbb{Z}_{p^n} is local and $J(\mathbb{Z}_{p^n}) = M$.

EXERCISES 3.4 HOMOMORPHISMS

1. (a) No (c) No (e) Yes
3. If $\theta : \mathbb{Z} \rightarrow \mathbb{Z}$ is a general ring homomorphism, let $\theta(1) = e$. Show that $e^2 = e$ so either $e = 1$ or $e = 0$. In the last case $\theta(k) = \theta(k \cdot 1) = \theta(k) \cdot \theta(1)$.
9. 0 and R up to isomorphism
10. (4) Clearly, $\theta(r^0) = \theta(1) = 1 = \theta(r)^0$. If $\theta(r^n) = \theta(r)^n$ for some $n \geq 0$, then $\theta(r^{n+1}) = \theta(r^n \cdot r) = \theta(r^n) \cdot \theta(r) = \theta(r)^n \cdot \theta(r) = \theta(r)^{n+1}$.
 (5) Note first that $\theta(u) \cdot \theta(u^{-1}) = \theta(uu^{-1}) = \theta(1) = 1$ and, similarly, that $\theta(u^{-1}) \cdot \theta(u) = 1$. So $\theta(u^{-1}) = \theta(u)^{-1}$. If $k \geq 0$, then (4) gives (5): If $k = -m$, $m > 0$, then $\theta(u^k) = \theta([u^{-1}]^m) = \theta(u^{-1})^m = [\theta(u)^{-1}]^m = \theta(u)^k$.
13. In \mathbb{Z}_7 this is $4n^2 = 2$ and this has a solution ($n = 2$) in \mathbb{Z}_7 . In \mathbb{Z}_{11} it is $7m^2 = 9$, or $m^2 = 8 \cdot 9 = 72 = 6$. But $m^2 = 0, 1, 3, 4, 5, 9$ in \mathbb{Z}_{11} , so there is no solution.
17. $R \cong R$ for any ring R because $1_R : R \rightarrow R$ is an isomorphism. If $R \cong S$, say $\sigma : R \rightarrow S$ is an isomorphism, then $\sigma^{-1} : S \rightarrow R$ is also an isomorphism, so $S \cong R$. If also $S \cong T$, where $\tau : S \rightarrow T$ is an isomorphism, then $\tau\sigma : R \rightarrow T$ is an isomorphism and $R \cong T$.
21. If $\theta : \mathbb{C} \rightarrow \mathbb{R}$ is a ring homomorphism, then $\ker \theta \neq \mathbb{C}$ because $1 \notin \ker \theta$ ($\theta(1) = 1 \neq 0$). Thus $\ker \theta = 0$ because \mathbb{C} is a field, from which $\mathbb{C} \cong \theta(\mathbb{C}) \subseteq \mathbb{R}$. If $\theta(i) = a$, then $a^2 = \theta(i)^2 = \theta(i^2) = \theta(-1) = -1$, a contradiction.
29. (a) The map $\theta : R(w) \rightarrow \frac{R}{A}(w)$ given by $\theta(a + bw) = \bar{a} + \bar{b}w$ is an onto homomorphism with kernel $A(w)$.
37. Define $\theta : \mathbb{Z} \rightarrow \mathbb{Z}_m \times \mathbb{Z}_n$ by $\theta(k) = (k + m\mathbb{Z}, k + n\mathbb{Z})$. Show that θ is a ring homomorphism and $\ker(\theta) = t\mathbb{Z}$.
41. If $\theta = \frac{1}{2}(1 + u\sqrt{2})$, show that $e^2 = e$. If $\sigma : R \rightarrow R(\sqrt{2})e$ is defined by $\sigma(r) = re$, show that σ is a ring isomorphism. Hence $R \cong R(\sqrt{2})e$.

EXERCISES 3.5 ORDERED INTEGRAL DOMAINS

3. (a) If $a \geq 0$, then $|a| = a \geq 0$. If $a < 0$, then $-a = 0 - a \in R^+$, so $|a| = -a > 0$.
 (c) If $a = 0$ or $b = 0$, then $ab = 0$ and $|ab| = 0 = |a||b|$. Assume that $a \neq 0$ and $b \neq 0$.
 - (1) If $a > 0$ and $b > 0$, then $ab > 0$, so $|ab| = ab = |a||b|$.
 - (2) If $a > 0$ and $b < 0$, then $ab < 0$, so $|ab| = -ab = a(-b) = |a||b|$.

- (3) If $a < 0$ and $b > 0$, the argument is like (2).
 (4) If $a < 0$ and $b < 0$, then $ab > 0$, so $|ab| = ab = (-a)(-b) = |a||b|$.
 Hence, $|ab| = |a||b|$ in every case.

EXERCISES 4.1 POLYNOMIALS

3. (a) 500
 4. (a) In \mathbb{Z}_6 : 4, 5, 1, 2; in \mathbb{Z}_7 : 4, 5
 5. (a) In \mathbb{Z}_4 : 0, 1; in $\mathbb{Z}_2 \times \mathbb{Z}_2$ all four elements are roots; in \mathbb{Z}_6 : 0, 1, 3, 4
 7. (a) Let ux^n and bx^m be the leading terms of f and g , where u is a unit. The leading term of fg is ubx^{n+m} because $ub \neq 0$ (otherwise $b = u^{-1}(ub) = 0$). Hence, $fg \neq 0$ and $\deg(fg) = n + m = \deg f + \deg g$.
 14. (a) $q = x^3 + 3x^2 - 3x + 5$, $r = -x - 3 = 5x + 3$
 (c) $q = 3x^2 + 2x + 3$, $r(x) = 7$ (e) $q = 3x + 2$, $r(x) = -14x - 3$
 16. (a) 3, 5
 17. (a) $f = (x - 1)(x + 1)(x - 5)(x + 5)$
 (c) $f = (x - 1)(x + 2)(x + 3)$
 23. (a) 1 (c) 3
 25. (a) $\frac{3}{4}$ (c) 2, -1 (e) None
 34. (c) If $u \in \mathbb{C}$ let \bar{u} denote the conjugate of u . Define $\theta : \mathbb{C}[x] \rightarrow \mathbb{C}$ by $\theta[f(x)] = \overline{f(0)}$. This is a homomorphism (it is evaluation at 0 followed by conjugation) but it is not evaluation at a for any $a \in \mathbb{C}$. Indeed, if $\theta = \varphi_a$ then $\theta(i) = \bar{i} = -i$ while $\varphi_a(i) = i$.
 38. (a) Show that $R[x]/P[x] \cong (R/A)[x]$ as rings (Exercise 37). Use Theorem 2.

EXERCISES 4.2 FACTORIZATION OF POLYNOMIALS OVER A FIELD

4. (a) Irreducible (c) Not irreducible (e) Irreducible
 5. (a) Yes, no, no, no, yes, yes
 (c) Yes, yes, no, yes, no, yes, yes
 8. (a) As f is monic, we may assume that both factors are monic (Exercise 6). Hence $f = (x - u)(x - v) = x^2 - (u + v)x + uv$. Now equate coefficients.
 15. $x^4 + 2x^2 + 1$, $x^4 + x^3 + x + 2$, $x^4 + 2x^3 + 2x^2 + x + 1$, $x^4 + 2x^3 + 2x + 2$, $x^4 + x^3 + 2x^2 + 2x + 1$, $x^4 + 1$
 18. (a) $3x^4 + 2 = 3(x - 1)(x + 1)(x - 3)(x + 3)$ in $\mathbb{Z}_5[x]$.
 (c) $x^3 + 2x^2 + 2x + 1 = (x + 1)(x + 3)(x + 5)$ in $\mathbb{Z}_7[x]$.
 (e) $x^4 - x^2 + x - 1 = (x - 1)(x - 2)(x^2 + 3x + 6)$ in $\mathbb{Z}_{13}[x]$.
 22. (a) Eisenstein Criterion, with $p = 3$
 23. (a) $f(x + 1) = x^4 + 4x^3 + 6x^2 + 6x + 2$, so use the Eisenstein criterion with $p = 2$.
 31. (a) If f is irreducible in $K[x]$ it cannot factor properly in $F[x]$.
 35. (a) Already irreducible (c) $f = (x^2 + 3x - 1)(x^2 - x + 2)$
 39. (a) $1 = (4x^2 + 3x + 4)f - (4x + 2)g$
 (c) $x - 2 = \frac{1}{4}g - \frac{1}{4}(x^3 + x^2 - x - 1)f$
 42. (a) Let $1 = mf + kg$ with m and k in $F[x]$. If $h = pf$ and $h = qg$, then $h = hmf + hkg = (qg)mf + (pf)hg = (qm + pk)fg$.

EXERCISES 4.3 FACTOR RINGS OF POLYNOMIALS OVER A FIELD

2. (a)

$+$	0	1	t	$1+t$	\times	0	1	t	$1+t$
0	0	1	t	$1+t$	0	0	0	0	0
1	1	0	$1+t$	t	1	0	1	t	$1+t$
t	t	$1+t$	0	1	t	0	t	1	$1+t$
$1+t$	$1+t$	t	1	0	$1+t$	0	$1+t$	$1+t$	0

(c)

\times	0	1	t	t^2	$1+t$	$1+t^2$	$t+t^2$	$1+t+t^2$
0	0	0	0	0	0	0	0	0
1	0	1	t	t^2	$1+t$	$1+t^2$	$t+t^2$	$1+t+t^2$
t	0	t	t^2	1	$t+t^2$	$1+t$	$1+t^2$	$1+t+t^2$
t^2	0	t^2	1	t	$1+t^2$	$t+t^2$	$1+t$	$1+t+t^2$
$1+t$	0	$1+t$	$t+t^2$	$1+t^2$	$1+t^2$	$t+t^2$	$1+t$	0
$1+t^2$	0	$1+t^2$	$1+t$	$t+t^2$	$t+t^2$	$1+t$	$1+t^2$	0
$t+t^2$	0	$t+t^2$	$1+t^2$	$1+t$	$1+t$	$1+t^2$	$t+t^2$	0
$1+t+t^2$	0	$1+t+t^2$	$1+t+t^2$	$1+t+t^2$	0	0	0	$1+t+t^2$

(e)

\times	0	1	-1	t	$-t$	$1+t$	$1-t$	$-1+t$	$-1-t$
0	0	0	0	0	0	0	0	0	0
1	0	1	-1	t	$-t$	$1+t$	$1-t$	$-1+t$	$-1-t$
-1	0	-1	1	t	t	$-1-t$	$-1+t$	$1-t$	$1+t$
t	0	t	$-t$	0	0	t	t	$-t$	$-t$
$-t$	0	$-t$	t	0	0	$-t$	$-t$	t	t
$1+t$	0	$1+t$	$-1-t$	t	$-t$	$1-t$	1	-1	$-1+t$
$1-t$	0	$1-t$	$-1+t$	t	$-t$	1	$1+t$	$-1-t$	-1
$-1+t$	0	$-1+t$	$1-t$	$-t$	t	-1	$-1-t$	$1+t$	1
$-1-t$	0	$-1-t$	$1+t$	$-t$	t	$-1+t$	-1	1	$1-t$

3. $t^3 = 1+t$

\times	0	1	t	t^2	$1+t$	$1+t^2$	$t+t^2$	$1+t+t^2$
0	0	0	0	0	0	0	0	0
1	0	1	t	t^2	$1+t$	$1+t^2$	$t+t^2$	$1+t+t^2$
t	0	t	t^2	$1+t$	$t+t^2$	1	$1+t+t^2$	$1+t^2$
t^2	0	t^2	$1+t$	$t+t^2$	$1+t+t^2$	t	$1+t^2$	1
$1+t$	0	$1+t$	$t+t^2$	$1+t+t^2$	$1+t^2$	t^2	1	t
$1+t^2$	0	$1+t^2$	1	t	t^2	$1+t+t^2$	$1+t$	$t+t^2$
$t+t^2$	0	$t+t^2$	$1+t+t^2$	$1+t^2$	1	$1+t$	t	t^2
$1+t+t^2$	0	$1+t+t^2$	$1+t^2$	1	t	$t+t^2$	t^2	$1+t$

5. (a) $\mathbb{Z}_3[x]/\langle x^3 - x + 1 \rangle$ (c) $\mathbb{Z}_{11}[x]/\langle x^2 + x + 1 \rangle$ 6. (a) Here $t^2 = t$. Idempotents: 0, t , 1, $1-t$; nilpotents: 0; units: $a+bt$, where $a \neq 0 \neq a+b$ 7. (a) $5(-1+t+t^2)$

14. (a) $(x+t)(x+t^2)(x+t+t^2)$, where $t^3 = 1+t$
 (c) $(x-t)(x-1-t)(x+1-t)$, where $t^3 = t-1$

21. (a) If x^2+ax+b is not irreducible over F , it must have a root $u \in F$. Thus,
 $u^2+au+b=0$. Take $c=2u+a$. Then $c^2=4(u^2+ua)+a^2=-4b+a^2$, con-
 trary to hypothesis.

25. (a) Write polynomials as $f=f(x)$. Then $d \in \langle f \rangle + \langle g \rangle$ because $d=uf+vg$ for
 some $u,v \in F[x]$. On the other hand, $f \in \langle d \rangle$ and $g \in \langle d \rangle$ because d is a common
 divisor of f and g . Hence $\langle f \rangle \subseteq \langle d \rangle$ and $\langle g \rangle \subseteq \langle d \rangle$, so $\langle f \rangle + \langle g \rangle \subseteq \langle d \rangle$.

EXERCISES 4.4 PARTIAL FRACTIONS

$$2. \quad (a) \frac{1}{x} - \frac{2}{x^2 + x + 1}$$

$$(c) \frac{1}{x} - \frac{x}{x^2 + 1} + \frac{1-x}{(x^2+1)^2}$$

EXERCISES 4.5 SYMMETRIC POLYNOMIALS

2. (a) $(y^2z^2) + (x^3 + xyz + x^2z) + (x^2 + xz - yz) + (3x - 3y)$
 7. (a) $x_2^2x_3 < x_1x_3 < x_1x_2^2x_3 < x_1^2x_2$
 8. (a) $f = s_1s_2^2$ (c) $f = s_1s_2s_3 - 3s_2^2$
 11. $p_5 = s_1^5 - 5s_1^3s_2 + 5s_1^2s_3 + 5s_1s_2^2 - 5s_1s_4 - 5s_2s_3 + 5s_5$
 12. (a) $f = (n-1)s_1^2 - 2ns_2$
 13. (a) $x^3 - 17x^2 - 14x - 9$

EXERCISES 5.1 IRREDUCIBLES AND UNIQUE FACTORIZATION

7. ± 1

10. (a) Irreducible (c) Not irreducible
 12. (a) Irreducible (c) Not irreducible
 14. (a) Not irreducible (c) Irreducible .
 16. (a) If $p \sim q$, suppose p is irreducible. If $q = ab$ in R then $p \sim ab$ so $p \sim a$ or $p \sim b$. Thus $q \sim a$ or $q \sim b$, so q is irreducible. The converse is the same.
 23. No. $\mathbb{Z}(\sqrt{-5})$.
 27. Write $d = \gcd[a, \gcd(b, c)]$ and $d_1 = \gcd[\gcd(a, b), c]$. Then d divides a and $\gcd(b, c)$, so it divides all a, b , and c . Thus d divides $\gcd(a, b)$ and c , which gives $d|d_1$. Similarly $d_1|d$, so $d \sim d_1$. Moreover, this result shows that d divides a, b , and c and that every common divisor of a, b , and c divides d . Hence, $\gcd(a, b, c)$ exists and $d \sim \gcd(a, b, c)$.
 31. If $m \sim \text{lcm}(a_1, \dots, a_n)$ exists in R then $a_i|m$ for each i shows $\langle m \rangle \subseteq \langle a_i \rangle$ for each i , and hence that $\langle m \rangle \subseteq A$ where we write $A = \langle a_1 \rangle \cap \dots \cap \langle a_n \rangle$. But $r \in A$ means $a_i|r$ for each i , so $m|r$ by definition. Thus, $r \in \langle m \rangle$ and we have $\langle m \rangle = A$. Conversely, if $A = \langle m \rangle$ then $a_i|m$ for each i (because $m \in \langle a_i \rangle$); and, if $a_i|r$ for each i , then $r \in A = \langle m \rangle$ so $m|r$. Thus, m is a least common multiple of the a_i .
 35. Use Gauss' lemma.
 37. If $f = ug$, u is a unit, write $u = \frac{a}{b}$. Since f and g are primitive, show $a \sim b$.
 39. (a) Show that R is a subring of $\mathbb{Z}[x]$.
 40. (a) Show $\langle x \rangle \subset \langle \frac{1}{2}x \rangle \subset \langle \frac{1}{4}x \rangle \subset \dots$.

EXERCISES 5.2 PRINCIPAL IDEAL DOMAINS

1. No, $\mathbb{Z}[x]$ in $\mathbb{Q}[x]$.
5. Let $A = \langle a \rangle$, $a \neq 0$. If a is a unit then $|R/A| = 1$. Otherwise, by Theorem 4 §3.3, let B/A be any ideal of R/A , say $B = \langle b \rangle$. Show that there are at most finitely many such divisors b of a up to associates.
8. (a) Write $R = \mathbb{Z}_{(p)}$. R is a subring of \mathbb{Z} because $-\left(\frac{m}{n}\right) = \frac{-m}{n}$, $\frac{m}{n} \cdot \frac{m'}{n'} = \frac{mm'}{nn'}$ and $\frac{m}{n} + \frac{m'}{n'} = \frac{mn' + m'n}{nn'}$ and p does not divide nn' . Thus, R is an integral domain. Given $\frac{m}{n}$ in R , if p does not divide m then $\frac{m}{n}$ is a unit in R (with inverse $\frac{n}{m}$). Conversely, if $\frac{m}{n}$ is a unit, say $\frac{m}{n} \cdot \frac{m'}{n'} = 1$, then $mm' = nn'$ so p does not divide m (it does not divide n or n'), so $R^* = \{\frac{m}{n} \in R \mid p \text{ does not divide } m\}$.
13. (b) $a = (1 + \sqrt{-2})b + (-1 + \sqrt{-2})$, where $\delta(-1 + \sqrt{-2}) = 3 < 11 = \delta(b)$.
15. (b) $a = 5b + (-1)$, where $\delta(-1) = 1 < 2 = \delta(b)$
24. (a) $\mathbb{Z}(i)/A \cong \mathbb{Z}_2$
26. (a) (1) \Rightarrow (2). Given $a \neq 0$, $b \neq 0$, let $\langle a, b \rangle = \langle d \rangle$. Then $d = ra + sb$ for some r , $s \in R$, so if $k|a$ and $k|b$ in R then $k|d$. But $d|a$ and $d|b$ because $a, b \in \langle d \rangle$.
 (2) \Rightarrow (1). Given $A = \langle a, b \rangle$, clearly A is principal if $a = 0$ or $b = 0$. Otherwise let $g = \gcd(a, b) \sim d$ where $d = ra + sb$ for $r, s \in R$. Then $d \in A$ so $\langle d \rangle \subseteq A$. On the other hand, $d|a$ and $d|b$ so $a \in \langle d \rangle$ and $b \in \langle d \rangle$. Hence $\langle a, b \rangle \subseteq \langle d \rangle$.
35. No. If so, and $w = \sqrt{-2}$, then $-2 = w^2 > 0$. But $2 = 1 + 1 > 0$.

EXERCISES 6.1 VECTOR SPACES

1. (a) No (c) No
2. (a) Yes (c) No
7. (a) Dependent (c) Independent
11. $\{(1, -1, 0), (1, 1, 1), (a, 0, 0)\}$ for any $a \neq 0$ in \mathbb{R}
15. I, A, A^2, A^3, A^4 cannot be independent.
19. (a) $\{1, r, \dots, r^n\}$ is not independent because $\dim_F(R) = n$. So $a_0 + a_1r + \dots + a_nr^n = 0$ where the $a_i \in F$ are not all zero.
22. (b) If $\{u_1, \dots, u_m\} \subseteq \{u_1, \dots, u_m, \dots, u_n\}$ then $\sum_{i=1}^m a_i u_i = 0$ implies that $\sum_{i=1}^n a_i u_i = 0$ where $a_{m+1} = \dots = a_n = 0$. So $a_i = 0$ for $1 \leq i \leq m$.
25. If v_i is in $\text{span}\{v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n\}$, then $v_i = \sum_{j \neq i} a_j v_j$ so, writing $a_i = -1$, $\sum_{i=1}^n a_i v_i = 0$, with $a_i \neq 0$. Hence $\{v_1, \dots, v_n\}$ is dependent. Conversely, if $\sum_{i=1}^n a_i v_i = 0$, where some $a_i \neq 0$, then $v_i = \sum_{j \neq i} (-a_i^{-1} a_j) v_j$ is in $\text{span}\{v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_n\}$.
31. (a) They are additive subgroups by group theory. If $v \in \ker \varphi$ then $\varphi(av) = a\varphi(v) = a0 = 0$ for all $a \in F$. If $w \in \text{im} \varphi$, say $w = \varphi(v)$, then $aw = a\varphi(v) = \varphi(av) \in \text{im} \varphi$.

EXERCISES 6.2 ALGEBRAIC EXTENSIONS

1. (a) $u^4 - 16u^2 + 4 = 0$ (c) $u^8 + 2u^4 + 49 = 0$
2. (a) $x^4 - 10x^2 + 1$ (c) $x^4 - 2x^2 - 2$
3. (a) Algebraic (c) Transcendental
4. (a) $x^2 - 2x + 2$

7. (a) $(u-\sqrt{3})^2 = (-i)^2 = -1$ so $u^2 - 2\sqrt{3}u + 4 = 0$. We claim that the minimal polynomial is $m(x) = x^2 - 2\sqrt{3}x + 4 \in \mathbb{R}[x]$. Its roots in \mathbb{C} are $\sqrt{3} \pm i$ and neither is in \mathbb{R} , so it is irreducible in $\mathbb{R}[x]$, as required.
12. (a) $\{1, u, u^2\}$, where $u = \sqrt[3]{2}$ (c) $\{1, u, u^2, \sqrt{3}, \sqrt{3}u, \sqrt{3}u^2\}$, where $u = \sqrt[3]{3}$
(e) $\{1, \sqrt{3}, \sqrt{5}, \sqrt{15}\}$
13. (a) 2 (c) 2
17. If $F(u) \supseteq L \supseteq F$ write $p = [F(u) : F] = [F(u) : L][L : F]$. Thus $[L : F] = 1$ or p ; so $L = F$ or $[L : F] = [F(u) : F]$, whence $L = F(u)$ by Theorem 8 §6.1.
19. Let $u \in E - \mathbb{Q}$. Show that $f \in \mathbb{Q}(u)$ with degree 2 exists such that $f(u) = 0$, say $f = ax^2 + bx + c$. Conclude that $[\mathbb{Q}(u) : \mathbb{Q}] = 2$, so $E = \mathbb{Q}(u)$. We may assume a, b , and c are integers, so $E = \mathbb{Q}(\sqrt{d})$. If $d = p^2e$, $e \in \mathbb{Z}$, p a prime, then $E = \mathbb{Q}(p\sqrt{e}) = \mathbb{Q}(\sqrt{e})$. Continue until $E = \mathbb{Q}(\sqrt{m})$ where m is square free.
21. (a) Write $L = F(u)$, so $F(u, v) = L(v)$. Thus $L(v) \supseteq L \supseteq F$ and $[L : F] = m$ by hypothesis, and $[F(u, v) : F] = [L(v) : L] \cdot m$. Hence, we simply show that $[L(v) : L] \leq n$. If p and m are the minimal polynomials of v over L and F , respectively, then $p|m$ by Theorem 3, so $[L(v) : L] = \deg p \leq \deg m = n$.
23. If $\sqrt{2} \in \mathbb{Q}(\pi)$ then $\sqrt{2} = \frac{f(\pi)}{g(\pi)}$, $f, g \in \mathbb{Q}[x]$, $\gcd(f, g) = 1$. Then $h(\pi) = 0$, where $h(x) = 2g^2(x) - f^2(x)$, and this is a contradiction if $h \neq 0$ in $\mathbb{Q}[x]$. But $h = 0$ means $2g^2 = f^2$ so, since $\gcd(f, g) = 1$, $f|2$. Thus $f = \pm 1, \pm 2$, $g^2(x) = \pm 1, \pm \frac{1}{2}$. This forces $\deg g = 0$; $g \in \mathbb{Q}$, $g = \pm 1, \pm \frac{1}{\sqrt{2}}$. Thus $g = \pm 1$, $\sqrt{2} = \frac{f(\pi)}{g(\pi)} = \pm 1$, a contradiction. Thus $h \neq 0$ and $\sqrt{2} \in \mathbb{Q}(\pi)$ has led to a contradiction. So $\sqrt{2} \notin \mathbb{Q}(\pi)$.
26. (a) Let $f(u^2) = 0$, $0 \neq f \in F[x]$. Use g where $g(x) = f(x^2) \neq 0$.
31. (a) We show $F(u) = Q$, where $Q = \{f(u)g(u)^{-1} \mid f, g \in F[x]; g(u) \neq 0\}$. Since u is transcendental over F , $f(u) \neq 0$ whenever $f(x) \neq 0$. Thus Q is a subfield of E containing F and u , so $F(u) \subseteq Q$. But any subfield of E containing F and u must contain Q , so $F(u) \supseteq Q$. Thus $F(u) = Q$.
33. Show that if u is algebraic over A then $u \in A$, contrary to hypothesis. If $f(u) = 0$ where $f \neq 0$ in $A[x]$, let $f = w_0 + w_1x + \cdots + w_nx^n$, $w_i \in A$. Show that $u \in L(u)$ where $L = F(w_1, \dots, w_n)$ is a finite extension of F .

EXERCISES 6.3 SPLITTING FIELDS

1. (a) $E = \mathbb{Q}(i\sqrt{3})$, and $[E : Q] = 2$. (c) $E = \mathbb{Q}(i, \sqrt{7})$, and $[E : Q] = 4$.
2. (a) $\mathbb{Q}(\sqrt{3}, \sqrt{5})$
4. (a) $E = \mathbb{Z}_2(u)$, $u^2 + u + 1 = 0$; $f(x) = (x+1)(x+u)(x+1+u)$
(c) $E = \mathbb{Z}_2(u)$, $f(u) = 0$; $f(x) = (x+u)(x+u^2)(x+1+u+u^2)$
(e) $E = \mathbb{Z}_3(u)$, $u^2 + 1 = 0$, $f(x) = (x-u)^2(x+u)^2$
6. (a) No. If \mathbb{C} were the splitting field of $f(x) \in \mathbb{Q}[x]$ then $\mathbb{C} = \mathbb{Q}(u_1, \dots, u_n)$. Thus $\mathbb{C} \supseteq \mathbb{Q}$ would be algebraic, contradicting the fact that π or e is transcendental.
9. If $\gcd(f, g) = 1$ let $1 = fh + gk$; h, k in $F[x]$. Suppose $E \supseteq F$ is an extension containing a common root u , that is $f(u) = 0 = g(u)$. Then substitution gives $1 = f(u)h(u) + g(u)k(u) = 0$, a contradiction. So no such extension E exists. Conversely, let $d = \gcd(f, g)$. If $d \neq 1$ then $\deg d \geq 1$ so let $E \supseteq F$ be a field containing a root u of d . Then $d|f$ and $d|g$ means $f(u) = 0 = g(u)$, contrary to hypothesis. So $d = 1$.
13. Show that the roots of $x^p - 1$ are $1, w, w^2, \dots, w^{p-1}$, so $\mathbb{Q}(w)$ is the splitting field. By Theorem 6, Appendix A, $x^p - 1 = (x-1)\Phi_p$, where $\Phi_p = x^{p-1} + x^{p-2} + \cdots + x + 1$ is irreducible over \mathbb{Q} by Example 13 §4.2.

20. (a) We show $A = \mathbb{Q}(\pi)$. Clearly $A \supseteq \mathbb{Q}$ is algebraic. We must show that if $u \in E$ is algebraic over \mathbb{Q} then $u \in A$. Since u is algebraic over A , we show that $u \notin A$ implies u is transcendental over A . We have $E = A(\pi)$ so this follows from Exercise 31 §6.2 if we can show that π is transcendental over A . But if π were algebraic over A it would be algebraic over $\mathbb{A} \supseteq A$, contrary to the preceding exercise.

EXERCISES 6.4 FINITE FIELDS

1. (a) 2
 4. (a)

(c) Any element of $GF(8)$ except 0 and 1.

(c)

5. If $GF(16) = \{a + bt + ct^2 + dt^3 \mid a, b, c, d \text{ in } \mathbb{Z}_2, t^4 = t + 1\}$, then t is primitive. The subfields are $GF(2^4) = GF(16)$, $GF(2) = \mathbb{Z}_2$, and $GF(2^2) = \{0\} \cup \langle t^5 \rangle = \{0, 1, t^5, t^{10}\} = \{0, 1, t + t^2, 1 + t + t^2\}$.

8. $\frac{n}{m}$.

9. If $G \subseteq C^*$, $|G| = n$, then $G = \langle u \rangle$, where $u = e^{2\pi i/n}$.

17. Let $d = \gcd(f, f')$ and write $d = fg + f'h$ where g and h are in $F[x]$. If $d = 1$, suppose f has a repeated root a in $E \supseteq F$. Then $x - a$ divides both f and f' in $E[x]$, and so divides $d = 1$, a contradiction. Conversely, if $d \neq 1$, let E be a splitting field of f over F . Then $d|f$ implies d has a root a in E , so $x - a$ divides f and f' , a contradiction by Theorem 3.

22. (a) If $K \supseteq \mathbb{Z}_p$ is a field containing a root u of f , conclude that f is the minimal polynomial of u over \mathbb{Z}_p . If $E = \mathbb{Z}_p(u)$ then $[E : \mathbb{Z}_p] = n$ and so $|E| = p^n$. Then u is a root of $h = x^{p^n} - x$, so $f|h$ in $E[x]$, say $h = qf$. But $h = q_0f + r$ in $\mathbb{Z}_p[x]$ by the division algorithm, so this holds in $E[x]$. By the uniqueness in $E[x]$, we get $q = q_0 \in \mathbb{Z}_p[x]$ and $r = 0 \in \mathbb{Z}_p[x]$.

EXERCISES 6.5 GEOMETRIC CONSTRUCTIONS

3. Yes. Bisect 30° , after constructing that from a $30-60-90$ -triangle.

5. No. A sphere of radius 1 has volume $\frac{4}{3}\pi$ and, if this is the volume of a cube with side a , then $a = \left(\frac{4\pi}{3}\right)^{1/3}$. But a is not constructible since it is not even algebraic over \mathbb{Q} . For if a is a root of $f(x) \in \mathbb{Q}[x]$. Then π is a root of $g(x) = f[\frac{3}{4}x^3]$. This is impossible as π is transcendental over \mathbb{Q} .

EXERCISES 6.7 AN APPLICATION TO CYCLIC AND BCH CODES

5. (a) In B_4 : $1+t$, $t+t^2 = t(1+t)$, $t^2+t^3 = t^2(1+t)$ and $1+t^3 = t^3(1+t)$. The other members 0 , $1+t^2$, $t+t^2$ and $1+t+t^2+t^3$ all lie in smaller ideals. (See Example 3.)

7. (a) $1+x^7 = (1+x)(1+x+x^3)(1+x^2+x^3)$ so there are $2 \cdot 2 \cdot 2 = 8$ divisors in all. Thus, there are 7 codes (excluding $\langle 1+x^7 \rangle = 0$).
11. Since $g(x) = 1+x+x^4$ has no root in \mathbb{Z}_2 , if it factorizes at all, it must do so as $g(x) = (a+bx+cx^2)(a'+b'x+c'x^2)$. Thus $aa'=1=cc'$ so $a=a'=1=c=c'$. Thus $g(x) = (1+bx+x^2)(1+b'x+x^2)$ so (coefficient of x^3) $b+b'=0$ and (coefficient of x) $b+b'=1$. This is impossible.
17. (c) Write $g(x) = 1+x+x^3$. We have $1+x^7 = (1+x)(1+x^2+x^3)g(x)$, a product of irreducibles. Hence, $1+x^7 = h(x)g(x)$, where $h(x) = 1+x+x^2+x^4$. We have $1 = xg(x) + 1 \cdot h(x)$, so take $e(x) = xg(x) = x+x^2+x^4$. Note that $e(t)^2 = e(t^2) = t^2+t^4+t^8 = t^2+t^4+t = e(t)$. So $e(t) = t+t^2+t^4$ is the idempotent generator.

EXERCISES 7.1 MODULES

1. (c) Using (a), $x + (-x) = 0 = 0x = (1 + (-1))x = 1x + (-1)x = x + (-1)x$, so $-x = (-1)x$.
2. (a) If $\alpha : M \rightarrow N$ is onto and R -linear, and if $M = Rx_1 + \cdots + Rx_n$, then we have $N = R\alpha(x_1) + \cdots + R\alpha(x_n)$. Since some of the $\alpha(x_i)$ may be zero, the result follows.
5. (a) If $x = \sum a_i k_i$ then $rx = \sum (ra_i)k_i \in AK$
6. Let $A = \sum_i Ra_i$, $a_i \in A$, and $M = \sum_j Rx_j$, $x_j \in M$. Use Exercise 5 to show that $AM = \sum_{i,j} Ra_i x_j$.
7. (a) Define $\alpha : N \rightarrow \frac{K+N}{K}$ by $\alpha(n) = n + K$ for all $n \in N$. Show that α is R -linear and $\ker \alpha = N \cap K$. Every coset in $\frac{K+N}{K}$ has the form $(k+n) + K$, where $k \in K$ and $n \in N$. But $(k+n) + K = n + K = \alpha(n)$, which proves that α is onto. Now the isomorphism theorem applies.
11. (a) Yes. $(m, n) = (n, n) + (m - n, 0)$ shows that $M = K + X$; clearly $K \cap X = 0$.
16. (a) We have $M = \pi(M) + \ker \pi$ because $m = \pi(m) + (m - \pi(m))$ for each $m \in M$ and $\pi[m - \pi(m)] = \pi(m) - \pi^2(m) = 0$. If $m \in \pi(M) \cap \ker \pi$, let $m = \pi(m_1)$ with $m_1 \in M$. Then $0 = \pi(m) = \pi^2(m_1) = \pi(m_1) = m$, so $\pi(M) \cap \ker \pi = 0$.
21. (a) If $w + AW = w_1 + AW$; we must show that $\alpha(w) + AV = \alpha(w_1) + AV$. Show that $w - w_1 = \sum_i a_i w_i$ where $a_i \in A$, $w_i \in W$, and apply the linearity of α .

EXERCISES 7.2 MODULES OVER A PID

1. (c) $\mathbb{Z}_4 \oplus \mathbb{Z}_3$, $\mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3$.
2. (a) The types are (4) , $(3, 1)$, $(2, 2)$, $(2, 1, 1)$ and $(1, 1, 1, 1)$. Hence, representative groups are \mathbb{Z}_{p^4} , $\mathbb{Z}_{p^3} \oplus \mathbb{Z}_p$, $\mathbb{Z}_{p^2} \oplus \mathbb{Z}_{p^2}$, $\mathbb{Z}_{p^2} \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$ and $\mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p \oplus \mathbb{Z}_p$.
3. (a) $\mathbb{Z}_p \oplus \mathbb{Z}_{q^2}$, $\mathbb{Z}_p \oplus \mathbb{Z}_q \oplus \mathbb{Z}_q$.
4. (a) The types are: p -component (2) , $(1, 1)$; the q -component (3) , $(2, 1)$, $(1, 1, 1)$; and the r -component (4) , $(3, 1)$, $(2, 2)$, $(2, 1, 1)$, $(1, 1, 1, 1)$. Hence $2 \cdot 3 \cdot 5 = 30$ in all.
7. (a) Thus $G(2)$ has type $(2, 2)$; $G(3)$ has type $(1, 1, 1)$ and $G(5)$ has type $(2, 1)$.
9. (a) The types are $(2, 2, 2)$, $(2, 2, 1, 1)$, $(2, 1, 1, 1, 1)$, $(1, 1, 1, 1, 1, 1)$.
12. (a) $T(K) = \{k \in K \mid o(k) \neq 0\} = K \cap \{m \in M \mid o(m) \neq 0\} = K \cap T(M)$.

16. (a) Define $\sigma : K \rightarrow M/T(M)$ by $\sigma(k) = k + T(M)$. This is a group homomorphism and $\ker \sigma = \{k \in K \mid k \in T(M)\} = K \cap T(M) = T(K)$. Use the isomorphism theorem.
20. (c) If $m = \sum x_i$ then $dm = 0$ implies $dx_i = 0$ for all i . Hence $L_d(M) \subseteq \sum_i L_d(M_i)$.
22. (a) Here $L_p(Rx) = \{rx \mid p(rx) = 0\}$. If $prx = 0$ then $pr \in \text{ann}(x) = \langle p^m \rangle = Rp^m$, say $pr = sp^m$. Since $m \geq 1$ and R is a domain, this gives $r = sp^{m-1}$, and we have shown that $(Rx)^p \subseteq R(p^{m-1}x)$. The other inclusion is because $p^m x = 0$. Finally, $p(Rx) = Rp x$ is a routine verification, and $px = 0$ if $m = 1$ because $o(x) = p^m$.
27. (a) $L_p(G)$ consists of 0 and the elements of order p . We have $L_p(G) = L_p(G_1) \oplus \cdots \oplus L_p(G_p)$ by Exercise 20, and $|L_p(G_i)| = p$ for each i by Exercise 26.

EXERCISES 8.1 FACTORS AND PRODUCTS

1. (a) $XY = \{\varepsilon, \sigma, \sigma^2\}$. Note: XY is a subgroup here, but X and Y are not.
3. $\frac{H}{G'} \triangleleft \frac{G}{G'}$ as $\frac{G}{G'}$ is abelian. Hence $H \triangleleft G$ by the correspondence theorem.
4. (a) $K, G, \langle a \rangle, \{1, a^3, b, ba^3\}, \{1, a^3, ba, ba^4\}, \{1, a^3, ba^2, ba^5\}$.
 (c) K, A_4
5. (a) $p\mathbb{Z}$, where p is any prime
 (c) If $D_{10} = \{1, a, \dots, a^9, b, ba, \dots, ba^9\}$, where $|a| = 10$, $|b| = 2$, and $aba = b$, the maximal subgroups are

$$\begin{aligned} H_1 &= \langle a \rangle \\ H_2 &= \langle a^2, b \rangle = \{1, a^2, a^4, a^6, a^8, b, ba^2, ba^4, ba^6, ba^8\} \\ K_0 &= \langle a^5, b \rangle = \{1, a^5, b, ba^5\} \\ K_1 &= \langle a^5, ba \rangle = \{1, a^5, ba, ba^6\} \\ K_2 &= \langle a^5, ba^2 \rangle = \{1, a^5, ba^2, ba^7\} \\ K_3 &= \langle a^5, ba^3 \rangle = \{1, a^5, ba^3, ba^8\} \\ K_4 &= \langle a^5, ba^4 \rangle = \{1, a^5, ba^4, ba^9\} \end{aligned}$$

9. (a) Clearly $\frac{H \cap H_1}{K} \subseteq \frac{H}{K} \cap \frac{H_1}{K}$. If $Kg \in \frac{H}{K} \cap \frac{H_1}{K}$, let $Kg = Kh$, $h \in H$. Then $gh^{-1} \in K \subseteq H$, so $g \in H$. Similarly $g \in H_1$, so $Kg \in \frac{H \cap H_1}{K}$.

12. (a) $H^2 \subseteq H$ because H is closed; $H \subseteq H^2$ because $1 \in H$.
15. $KA = AK$ and $KB = BK$ are subgroups by Theorem 5 §2.8. Given kb in KB , $Ab = bA$ and $Kb = bK$ by hypothesis, so $KA(kb) = AKkb = AKb = AbK = bAK = bKA = KbA = kKbA = (kb)KA$. Thus $KA \triangleleft KB$.
22. (a) Let $a \neq b$ have order 2, show that $H = \{1, a, b, ab\}$ is closed and apply Lagrange's theorem.

EXERCISES 8.2 CAUCHY'S THEOREM

1. (a) $\{1\}, \{a, a^3\}, \{a^2\}, \{b, ba^2\}, \{ba, ba^3\}$. The normal subgroups are the unions $\{1\}, \{1, a^2\}$ and $\{1, a, a^2, a^3\}, \{1, b, a^2, ba^2\}$ and $\{1, ba, a^2, ba^3\}$.
7. Let $K = g^{-1}Hg$, so $H = gKg^{-1}$. We claim $N(K) = g^{-1}N(H)g$. Let $a \in N(K)$ so $a^{-1}Ka = K$. To show $a \in g^{-1}N(H)g$ it suffices to show $gag^{-1} \in N(H)$. But $(gag^{-1})^{-1}H(gag^{-1}) = ga^{-1}g^{-1}Hgag^{-1} = ga^{-1}Kag^{-1} = gKg^{-1} = H$. Hence $N(K) \subseteq g^{-1}N(H)g$. A similar argument shows that $N(H) \subseteq gN(K)g^{-1}$ so $g^{-1}N(H)g \subseteq N(K)$.

11. Because $H \subseteq N(H) \subseteq G$, Exercise 31 §2.6 shows that $|G : N(H)|$ is finite. Hence, Theorem 2 applies.
14. We have $a^{-1}Ha = \{1, ba^2\}$ so $ba^2 \notin N(H)$. Continue in this way.
16. $N(\gamma) = \langle \gamma \rangle$.
25. H is a union of conjugacy classes. Show that there exists $a \neq 1$ such that $\{a\} \subseteq H$.
29. Since $C \triangleleft G$, let $Z[G/C] = K/C$. Since $|G/C| > 1$, Theorem 6 shows $C \subset K$. But $K \triangleleft G$ so $K \not\subseteq H$ by Exercise 23 §2.8. If $k \in K$ then kC is in the center of G/C , so $h^{-1}k^{-1}hk \in C \triangleleft H$. Hence $k^{-1}Hk \subseteq H$, and similarly $kHk^{-1} \subseteq H$. Thus $k \in N(H)$ and we have shown $K \subseteq N(H)$.

EXERCISES 8.3 GROUP ACTIONS

1. (a) By Cauchy's Theorem let $a \in G$, $|a| = 5$. If $H = \langle a \rangle$, then $|G \cdot H| = 4$, so there is a homomorphism $\theta : G \rightarrow S_4$ with $\ker \theta \subseteq H$. Then $\ker \theta \neq \{1\}$ because $|G| = 20$ does not divide $|S_4| = 24$. Because H is simple, $\ker \theta = H$, so $H \triangleleft G$.
10. (a) $H_0 \subseteq H$ because $H = 1_G(H)$. If $\tau \in \text{aut } G$ then $\tau^{-1}\sigma \in \text{aut } G$ for all $\sigma \in \text{aut } G$, so $H_0 \subseteq \tau^{-1}\sigma(H)$. Thus $\tau(H_0) \subseteq \sigma(H)$ for all σ , so $\tau(H_0) \subseteq H_0$. Similarly $\tau^{-1}(H_0) \subseteq H_0$, whence $\tau(H_0) = H_0$. Thus H_0 is characteristic in G .
15. If $\sigma = (k_1 \ k_2 \ \dots)(m_1 \ m_2 \ \dots)(n_1 \ n_2 \ \dots)\dots$, the orbits of the group G in X_n are $G \cdot k_1 = \{k_1, k_2, \dots\}$, $G \cdot m_1 = \{m_1, m_2, \dots\}$, $G \cdot n_1 = \{n_1, n_2, \dots\}, \dots$. Clearly, $G \cdot k = \{k\}$ if and only if k is fixed by σ .
17. (a) $x \equiv x$ because $x = x \cdot 1$; if $x \equiv y$, say $y = x \cdot a$, $a \in G$, then $x = y \cdot a^{-1}$, so $y \equiv x$; if $x \equiv y$ and $y \equiv z$, say $y = x \cdot a$, $z = y \cdot b$, then $z = (x \cdot a) \cdot b = x \cdot (ab)$, so $z \equiv x$.
23. (a) If $a, b \in S(x)$ then $(ab) \cdot x = a \cdot (b \cdot x) = a \cdot x = x$, so $ab \in S(G)$. Similarly, $a^{-1} \cdot x = a^{-1} \cdot (a \cdot x) = (a^{-1} \cdot a) \cdot x = 1 \cdot x = x$, so $a^{-1} \in S(G)$. Finally $1 \cdot x = x$ shows $1 \in S(x)$, and we are done.
28. If $X = \{H \subseteq G \mid H \text{ is a subgroup and } |H| = p^k\}$, let G act on X by conjugation. Use Theorem 4.
32. (a) $(1, 1) \cdot x = 1x1^{-1} = x$, and

$$(h_1, k_1) \cdot ((h, k) \cdot x) = h_1(hxk^{-1})k_1^{-1} = (h_1h, k_1k) \cdot x = [(h_1k_1) \cdot (h, k)] \cdot x.$$
The orbit is

$$(H \times K) \cdot x = \{(h, k) \cdot x \mid h \in H, k \in K\} = \{hxk^{-1} \mid h \in H, k \in K\} = HxK.$$

EXERCISES 8.4 THE SYLOW THEOREMS

1. If P is a Sylow 3-subgroup, then $P = \langle \gamma \rangle$, where γ is a 3-cycle, say $\gamma = (i \ j \ k)$. If $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ i & j & k & x \end{pmatrix}$, where $\{1, 2, 3, 4\} = \{i, j, k, x\}$, then $\sigma(1 \ 2 \ 3)\sigma^{-1} = \gamma$. Hence $\sigma \langle (1 \ 2 \ 3) \rangle \sigma^{-1} = P$, so P is conjugate to $\langle (1 \ 2 \ 3) \rangle$.
3. P is a Sylow p -subgroup of $N(P)$, being a p -subgroup of maximal order. It is unique because it is normal in $N(P)$.
7. (a) $|G| = 40 = 2^3 \cdot 5$. Thus, $n_5 = 1, 2, 4, 8$, and $n_5 \equiv 1 \pmod{5}$. Hence, $n_5 = 1$, so the Sylow 5-subgroup is normal.
(c) $|G| = 48 = 2^4 \cdot 3$. If P is a Sylow 2-subgroup then $|G : P| = 3$, so a homomorphism $\theta : G \rightarrow S_3$ exists. Clearly, $\ker \theta \neq \{1\}$.

9. (a) $|G| = 70 = 2 \cdot 5 \cdot 7$. Then $n_5 = 1, 2, 7, 14$, and $n_5 \equiv 1 \pmod{5}$, so $n_5 = 1$. Similarly $n_7 = 1$, so let $P \triangleleft G$ and $Q \triangleleft G$, where $|P| = 5$ and $|Q| = 7$. Because $P \cap Q = \{1\}$, $PQ \cong P \times Q \cong C_5 \times C_7 \cong C_{35}$. Hence, $|G : PQ| = 2$, so $PQ \triangleleft G$.
11. (a) $|G| = 105 = 3 \cdot 5 \cdot 7$. Then $n_7 = 1, 3, 5, 15$, and $n_7 \equiv 1 \pmod{7}$, so $n_7 = 1, 15$. Similarly, $n_5 = 1, 21$. Let P and Q by Sylow 7- and 5-subgroups. If neither is normal in G , then G has $21 \cdot 4 = 84$ elements of order 5 and $15 \cdot 6 = 90$ elements of order 7, a contradiction. So $P \triangleleft G$ or $Q \triangleleft G$; hence PQ is a subgroup, and $|PQ| = |P||Q| = 35$ because $P \cap Q = \{1\}$. As $|G : PQ| = 3$, let $\theta : G \rightarrow S_3$ be a homomorphism with $\ker \theta \subseteq PQ$. Then $|\ker \theta| \neq 1, 5, 7$, so $PQ = \ker \theta \triangleleft G$. Finally, $P \triangleleft PQ$ and $Q \triangleleft PQ$ by the Sylow Theorems, so $PQ \cong P \times Q \cong C_7 \times C_5 \cong C_{35}$.
13. Let P be a Sylow p -subgroup of G . Since $p > m$ we have $|P| = p^n$, so $|G : P| = m$. Apply Theorem 1 §8.3.
19. If Q is also a Sylow p -subgroup of G , then $Q = a^{-1}Pa$ by Sylow's second theorem. If $g \in N(Q)$ then $Q = g^{-1}Qg$; that is $a^{-1}Pa = g^{-1}a^{-1}Pag$. This implies that $aga^{-1} \in N(P) = P$, whence $g \in a^{-1}Pa = Q$.

EXERCISES 8.5 SEMIDIRECT PRODUCTS

1. (a). Write $\sigma = (1\ 2) \in S_n$ and $H = \langle \sigma \rangle$. Then $A_n \subseteq A_nH \subseteq S_n$. As $S_n/A_n \cong C_2$, either $A_nH = S_n$ or $A_nH = A_n$. Since $\sigma \notin A_n$ we have $S_n = A_nH$. Similarly, $A_n \cap H \neq \{\varepsilon\}$ means $A_n \cap H = H$ (because H is simple), contradicting $h \notin A_n$ once more. Hence $A_n \cap H = \{\varepsilon\}$ and the result follows from Theorem 2.
3. This is an instance of Theorem 3 (3), where $p = 3$ and $q = 13$. We have $q \equiv 1 \pmod{p}$, so we look for m such that $1 \leq m \leq 12$ and $m^5 \equiv 1 \pmod{13}$. If $m = 1$ then $G \cong C_{13} \times C_3 \cong C_{55}$. The first solution with $m > 1$ is $m = 3$, whence $G \langle a, b \rangle$ where $o(a) = 11$, $o(b) = 3$ and $ab = ba^3$.

EXERCISES 8.6 AN APPLICATION TO COMBINATORICS

6. (a) $\frac{1}{12}q^2(q^2 + 11)$
 8. (a) $\frac{1}{6}q(q + 1)(q^4 - q^3 + q^2 + 2)$

EXERCISES 9.1 THE JORDAN–HÖLDER THEOREM

1. (a) 3; C_2, C_2, C_2 (c) 3; C_2, C_2, C_2 (e) 3; C_2, C_2, C_2
 3. (a) If H_k is the unique subgroup of order k in C_{24} , the series are

$$\begin{aligned} C_{24} &\supset H_{12} \supset H_4 \supset H_2 \supset \{1\} \\ C_{24} &\supset H_{12} \supset H_6 \supset H_3 \supset \{1\} \\ C_{24} &\supset H_{12} \supset H_6 \supset H_2 \supset \{1\} \\ C_{24} &\supset H_8 \supset H_4 \supset H_2 \supset \{1\} \end{aligned}$$

8. (a) Let $n = p_1p_2 \cdots p_m$, where the p_i are distinct primes. Then C_n has length $1 + 1 + \cdots + 1 = m$ by Example 8.
11. Induct on n . If $n = 1$ then $G = G_0 \supset G_1 = \{1\}$ so $G \cong G_0/G_1$ is finite. In general, G_1 is finite by induction, and $G/G_1 \cong G_0/G_1$ is finite by hypothesis. Thus G consists of $|G/G_1|$ cosets, each with $|G_1|$ elements. Hence G is finite. Now $|G| = |G_0/G_1| \cdot |G_1|$, and the formula follows by induction.

15. (a) If $M \subseteq C_n$ is maximal normal, then C_n/M has order a prime q (being simple and abelian). Hence, $q = p_i$ for some i because q divides $|C_n| = n$. Thus, $|M| = \frac{n}{p_i}$ for some $i = 1, 2, \dots, r$. Since C_n is cyclic, it has exactly one subgroup of order $\frac{n}{p_i}$ by Theorem 9 §2.4.

EXERCISES 9.2 SOLVABLE GROUPS

1. No, $Z(S_4) = \{\varepsilon\}$.
3. No. S_4 is solvable (Example 4) but $S'_4 = A_4$ is not abelian. Indeed $S'_4 \subseteq A_4$ because S_4/A_4 is abelian. Thus $S'_4 = A_4$, $\{\varepsilon\}$ or $K = \{\varepsilon, (1 2)(3 4), (1 3)(2 4), (1 4)(2 3)\}$. But $S_4/\{\varepsilon\}$ and S_4/K are not abelian (see Exercise 30 §2.9).
8. (a) This is because $\alpha[a, b] = [\alpha(a), \alpha(b)]$ for every commutator $[a, b]$ from G .
9. By Exercise 14 §8.4, let $K \triangleleft G$, $K \neq \{1\}, G$. Then both $|K|$ and $|G/K|$ are in $\{p, q, p^2, pq\}$. Hence, both are either abelian or of order pq and thus are solvable. Use Theorem 4.
15. If G is solvable and $G = G_0 \supset G_1 \supset \dots \supset G_p = \{1\}$ is a composition series, each simple factor is abelian and hence finite. Hence, $|G| = \left|\frac{G_0}{G_1}\right| \cdot \left|\frac{G_1}{G_2}\right| \cdots \left|\frac{G_{n-1}}{G_n}\right|$ is finite (see Exercise 11 §9.1). The converse holds because every finite group has a composition series.
19. HK is a subgroup as $K \triangleleft G$, and $\frac{HK}{K} \cong \frac{H}{H \cap K}$ is solvable by Theorem 3 (H is solvable). Done by Theorem 4.
21. (a) Because $G \neq \{1\}$, $G' \neq G$ by Theorem 5. Thus G/G' is nontrivial and abelian.
23. (a) Write $\{K \triangleleft G \mid G/K \text{ solvable}\} = \{K_1, K_2, \dots, K_m\}$. This set is nonempty as it contains G . Then $R = \bigcap_{i=1}^m K_i$ is normal and G/R is solvable by Exercise 18. If $K \triangleleft G$ and G/K solvable, then $R \subseteq K$ by definition.
27. (a) Write $V = \mathcal{V}(G)$. Then $V \triangleleft G$ because the intersection of normal subgroups is normal. Note that the intersection is not empty because $G \triangleleft G$ and G/G is in \mathcal{V} . If $V = K_1 \cap K_2 \cap \dots \cap K_n$ then G/V embeds in $\frac{G}{K_1} \times \dots \times \frac{G}{K_n}$ (as in Exercise 18) and $\frac{G}{K_1} \times \dots \times \frac{G}{K_n}$ is in \mathcal{V} by induction because \mathcal{V} is closed under taking direct products. Hence G/V is in \mathcal{V} , being isomorphic to a subgroup of a group in \mathcal{V} .

EXERCISES 9.3 NILPOTENT GROUPS

2. If $H \triangleleft G$ and $K \triangleleft G$ then $a^{-1}[h, k]a = [a^{-1}ha, a^{-1}ka] \in [H, K]$ for all $h \in H$ and $k \in K$.
6. (a) By induction on n , it suffices to show that $\Gamma_i(G \times H) \subseteq \Gamma_i(G) \times \Gamma_i(H)$. Do so by induction on i . If $i = 0$, then $\Gamma_0(G \times H) = G \times H = \Gamma_0(G) \times \Gamma_0(H)$. If the relation holds for $i \geq 0$, then $\Gamma_{i+1}(G \times H) = [\Gamma_i(G \times H), G \times H] \subseteq [\Gamma_i(G) \times \Gamma_i(H), G \times H]$, so it suffices to show that, if $A \subseteq G$, $B \subseteq H$, then $[A \times B, G \times H] \subseteq [A, G] \times [B, H]$. This outcome follows because $[(a, b), (g, h)] = ([a, g], [b, h])$.
9. If $n = 2^k$ then $|D_n| = 2^{k+1}$ so D_n is nilpotent by Example 3. Conversely, suppose $n = 2^k m$, $m > 1$ odd. Show that $\langle a^{2^k}, b \rangle \cong D_m$ so D_m is nilpotent by Theorem 1. But in this case $\{1, b\}$ is a Sylow 2-subgroup that is not normal, contradicting Theorem 4.
13. $K \cap Z(G) \neq \{1\}$ by Exercise 11. Thus, $K \cap Z(G) = K$ by the condition on K . Now every subgroup of K is normal in G , so $|K|$ is prime.

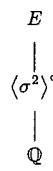
18. (a) H is itself nilpotent by Theorem 1 so, by Theorem 4, H is a product of p -groups. Now apply Theorem 6 §8.2.
21. (a) If $H = \langle a^2 \rangle$ then H and \bar{H} are maximal.
(c) If $H = \langle a^p \rangle$ and $H = \langle a^q \rangle$ then \bar{H} and \bar{K} are maximal, as is $\langle a^2 \rangle$.
24. (1) \Rightarrow (2). We show that $G' \subseteq \Phi$, that is $G' \subseteq M$ for every maximal subgroup M of G . Show that this follows by (1) because $M \triangleleft G$.
26. (a) Write $\Phi(G) = \Phi$ and $\Phi\left(\frac{G}{K}\right) = \frac{F}{K}$, where $F \triangleleft G$. If M is maximal in G then $K \subseteq M$ (because $K \subseteq \Phi \subseteq M$) and $\frac{M}{K}$ is maximal in $\frac{G}{K}$. Hence $\frac{F}{K} \subseteq \frac{M}{K}$ whence $F \subseteq M$. It follows that $F \subseteq \Phi$. Conversely, let $\alpha : G \rightarrow G/K$ be the coset map. Then $\alpha(\Phi) \subseteq \Phi(G/K)$ by the preceding exercise, so $x \in \Phi$ implies $xK \in \frac{F}{K}$; so $x \in F$. Thus $\Phi \subseteq F$.

EXERCISES 10.1 GALOIS GROUPS AND SEPARABILITY

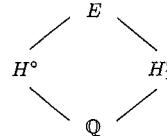
1. ε, σ^{-1} and $\sigma\tau$ all fix F when σ and τ do.
3. $\sigma(\sum a_i v_i) = \sum a_i \sigma(v_i)$ because $\sigma(a_i) = a_i$ for $a_i \in F$.
5. If $\sigma : E \rightarrow E$ is an automorphism, show that $\sigma(q) = q$ for all $q \in \mathbb{Q}$.
7. C_2
9. $C_2 \times C_2$
11. Show that $E = F(u)$ if $u \in E \setminus F$. If m is the minimal polynomial of u over F , show $\deg m = 2$.
13. Construct σ and τ in $\text{gal}(E : \mathbb{Q})$ with $\sigma(u) = iu$, $\sigma(i) = i$, and $\tau(u) = u$, $\tau(i) = -i$.
15. (a) $v = \sqrt{3}$ and $w = \sqrt{5}$ in Theorem 6.
16. $v = \sqrt{p}$ and $w = \sqrt{q}$ in Theorem 6.
17. See the Hint.
19. We proceed by induction on n . If $n = 1$ then $E = F(u_1) = \{f(u_1) \mid f(x) \in F[x]\}$. Hence $\sigma(f(u_1)) = g(\sigma(u_1)) = g(\tau(u_1)) = \tau(f(u_1))$ for all f , as required. In general, write $K = F(u_1, u_2, \dots, u_{n-1})$ so that $E = K(u_n)$. By induction, $\sigma = \tau$ on K , so $\sigma, \tau \in \text{gal}(K : F)$. Since $\sigma(u_n) = \tau(u_n)$ the result follows from the case $n = 1$.
21. See the Hint.
22. (a) For (3) \Rightarrow (1), if f has a repeated root u in $E \supseteq F$, let $1 = fg + f'h$ in $F[x]$ by (3).
23. If $d = \gcd(f, f')$, show $d = 1$.
25. If $E \supseteq F$ and q is an irreducible factor of f in $E[x]$, write $f = p_1 p_2 \cdots p_r$ in $F[x]$, p_i irreducible, and show that $q|p_i$ for some i .
27. If not, and u is a root of f in a splitting field $E \supseteq F$, show $f = (x - u)^p$ in $E[x]$. If q is an irreducible factor of f in $F[x]$, show $q = (x - u)^t$.
29. (a) If F is perfect, and $a \in F$, let E be the splitting field of $f = x^p - a$. If $u \in E$ is a root of f show $f = (x - u)^p$. If q is an irreducible factor of f in $F[x]$ show that $q = x - u$. Use Theorem 4.
30. (a) Let q be the minimal polynomial of u over F . If $K = F(u^p)$ let $m \in K[x]$ be the minimal polynomial of u over K . Then $q \in K[x]$ and $q(u) = 0$, so $m|q$. But q has distinct roots by hypothesis, so m has distinct roots. On the other hand, $x^p - u^p \in K[x]$ and $x^p - u^p = (x - u)^p$ in $E[x]$. Hence $m|(x - u)^p$ so $m = (x - u)^r$. Since m has distinct roots, $r = 1$ and so $u \in K$.
32. (a) Let p and q be the minimal polynomials of u over F and K respectively. Then $p \in K[x]$ and $p(u) = 0$, so $q|p$. Since p has distinct roots in some splitting field $L \supseteq K$, q is separable over K .

EXERCISES 10.2 THE MAIN THEOREM OF GALOIS THEORY

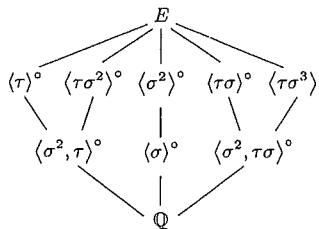
1. (a) By Example 4 §10.1, $\text{gal}(E, \mathbb{Q}) = \langle \sigma \rangle \cong C_4$, where $\sigma(u) = u^2$. If $H = \langle \sigma^2 \rangle$ then H° is the only intermediate field (except \mathbb{Q}, E). $H^\circ = \mathbb{Q}(u + u^4)$.



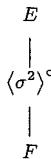
- (c) By Exercise 9 §10.1, $\text{gal}(E, \mathbb{Q}) = \langle \sigma, \tau \rangle \cong C_2 \times C_2$, where $\sigma(i) = -i$, $\sigma(\sqrt{3}) = \sqrt{3}$; and $\tau(i) = i$, $\tau(\sqrt{3}) = -\sqrt{3}$. If $H = \langle \sigma \rangle$ and $H_1 = \langle \tau \rangle$, the lattice of fields is as shown. $H^\circ = \mathbb{Q}(\sqrt{3})$; $H_1^\circ = \mathbb{Q}(i)$.



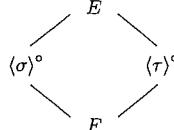
- (e) If $u = \sqrt[4]{2}$, then $E = \mathbb{Q}(u, i)$ and, by Exercise 13 §10.1, $\text{gal}(E : \mathbb{Q}) = \langle \sigma, \tau \rangle \cong D_4$, where $\sigma(u) = iu$, $\sigma(i) = i$; and $\tau(u) = u$, $\tau(i) = -i$. The lattice diagram is shown. Primitive elements are $E = \mathbb{Q}(u + i)$, $\langle \sigma^2 \rangle^\circ = \mathbb{Q}(u^2 + i)$, $\langle \tau \rangle^\circ = \mathbb{Q}(u)$, $\langle \sigma \rangle^\circ = \mathbb{Q}(i)$, $\langle \tau\sigma \rangle^\circ = \mathbb{Q}(u - iu)$, $\langle \sigma^2, \tau \rangle^\circ = \mathbb{Q}(u^2)$, $\langle \tau\sigma^2 \rangle^\circ = \mathbb{Q}(iu)$, $\langle \sigma^2, \tau\sigma \rangle^\circ = \mathbb{Q}(iu^2)$, and $\langle \tau\sigma^3 \rangle^\circ = \mathbb{Q}(u + iu)$.



2. (a) Either $G \cong C_{p^2} = \langle \sigma \rangle$:



- or $G \cong C_p \times C_p \cong \langle \sigma, \tau \rangle$:



5. (a) Let $r = r(t) = \frac{f(t)}{g(t)} \in E$ show $f(t)g(-t) = f(-t)g(t)$. If $\text{char } F \neq 2$, write $h(t) = f(t)g(-t)$. Show that $h(t) = k(t^2)$ for some polynomial k . Similarly and $g(t)g(-t) = l(t^2)$ for some polynomial l . Continue.

7. Clear.

9. (a) An intermediate field K is closed if $K = K'^\circ$.

11. Exercise 34 §2.6.

15. (a) Use the Galois connection.

17. If $K = \sigma(K_1)$ show $\sigma K'_1 \sigma^{-1} = K'$. If $\sigma^{-1} K'_1 \sigma = K$ show $K_1 \subseteq \sigma^{-1}(K)$, so $\sigma(K_1) \subseteq K$, and similarly $K \subseteq \sigma(K_1)$.

19. (a) Let $E = F(u_1, u_2, \dots, u_m)$ where u_1, \dots, u_m are the distinct roots of f , use Theorem 3 §10.1.

20. (a) Apply σ to the formulas for $N(u)$ and $T(u)$.

21. If $f = v_0 + v_1x + \cdots + v_mx^m$ show $f^\tau = \prod_{\sigma \in G} [x - \tau\sigma(u)] = f$.

EXERCISES 10.3 INSOLVABILITY OF POLYNOMIALS

1. (a) $\mathbb{Q}(\sqrt{3}, \sqrt[3]{5}, \sqrt[5]{7})$
2. (a) $f' = 5x^4 - 4$ has roots $\pm a$ and $\pm ia$, where $a = \sqrt[4]{4/5}$. Then $f(a) < 0$ and $f(-a) > 0$, so f has three real roots and two (conjugate) nonreal roots. As f is irreducible (Eisenstein), its Galois group is S_5 , as in Example 1.
3. Show that $p = x^7 - 14x + 2$ has three distinct real roots and two (conjugate) complex roots. If $E \supseteq \mathbb{Q}$ is the splitting field, view $G = \text{gal}[E : \mathbb{Q}]$ as a subgroup of S_X where $X \subseteq \mathbb{C}$ are the roots. Then conjugation is a transposition and, if u is a real root, then $[\mathbb{Q}(u) : \mathbb{Q}] = 7$ because p is the minimal polynomial of u over \mathbb{Q} . Proceed as in Example 1.
5. Let X denote the set of roots of p in a splitting field $E \supseteq F$ where $p \in F[x]$. View $G = \text{gal}(E : F) \subseteq S_X$, so G embeds in S_4 .
7. Since $f' = 3(x^2 - 1)$, conclude that f has three real roots. In the cubic formula, p^3 and q^3 are roots of $x^2 + x + 1$ which satisfy $p^3 + q^3 = -1$ and $pq = 1$. The roots are w and w^2 ($w = e^{2\pi i/3}$). Show that $p = e^{2\pi i/9}$ and $q = e^{16\pi i/9} = e^{-2\pi i/9} = \bar{p}$. The roots are $2 \cos\left(\frac{2\pi}{9}\right)$, $2 \cos\left(\frac{8\pi}{9}\right)$ and $2 \cos\left(\frac{4\pi}{9}\right)$. Finally $\text{gal}[E : \mathbb{Q}] \cong C_3$.
8. (a) $\sigma(\Delta^2) = \Delta^2$ for all $\sigma \in G$ because σ permutes the roots u_i .

EXERCISES 10.4 CYCLOTOMIC POLYNOMIALS AND WEDDERBURN'S THEOREM

1. (a) $x^4 + 1$ (c) $x^4 - x^2 + 1$ (e) $x^6 - x^3 + 1$
3. Use the Hint and induction.
5. If $w_k = e^{2\pi i/k}$, these fields are $\mathbb{Q}(w_{mn})$ and $\mathbb{Q}(w_m, w_n)$ respectively. Show that $\mathbb{Q}(w_m, w_n) = \mathbb{Q}(w_{mn})$. (\supseteq requires $\gcd(m, n) = 1$).
7. Write $\sigma(n) = \sum_{d|n} \mu(d)$. If $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$ and $m = p_1 p_2 \cdots p_r$, then $\sigma(n) = \sigma(m)$. If $d|m$, show $\mu(d) = 1$ if and only if d is the product of an even number (possibly 0) of the p_i , and $\mu(d) = -1$ otherwise.
8. (a)
$$\begin{aligned} \sum_{d|n} \mu(d) \alpha\left(\frac{n}{d}\right) &= \sum_{d|n} \mu(d) \left[\sum_{c|(n/d)} \beta(c) \right] = \sum_{cd|n} \mu(d) \beta(c) \\ &= \sum_{c|n} \beta(c) \left[\sum_{d|(n/c)} \mu(d) \right] = \beta(n) \end{aligned}$$
 by Exercise 7.

EXERCISES 11.1 WEDDERBURN'S THEOREM

1. Show that $M = Rx$ and use Theorem 1 §7.1.
3. Straight forward.
5. If $L \subseteq eRe$ is a left ideal, consider RL .
7. Use the Corollary to Lemma 2 and Theorem 5 §7.1.
8. (a) Show there exists a minimal member of $\mathcal{S} = \{\text{ann}(X) \mid X \subseteq M \text{ and } X \text{ is finite}\}$.

9. Let $X = \{\frac{m}{p^n} \mid m \in \mathbb{Z} \text{ and } p \text{ does not divide } m\}$. If $\mathbb{Z} \subset Y \subset X$, where Y is a subgroup of X , show that there exists $\frac{m}{p^n} \in Y$ with n maximal. Then show that $Y = \mathbb{Z} \frac{1}{p^n}$.
11. (a) If $x \in M$ show that $x - \pi(x) \in \ker \pi$. It follows that $M = \pi(M) + \ker \pi$. Continue.
13. (a) If $K = K_1 \oplus K_2 \oplus \dots$ then $K \supset K_2 \oplus K_3 \oplus \dots \supset K_3 \oplus K_4 \supset \dots$ and $K_1 \subset K_1 \oplus K_2 \subset K_1 \oplus K_2 \oplus K_3 \subset \dots$
15. Use the Hint.
17. (a) Show that $\ker(\alpha) \subseteq \ker(\alpha^2) \subseteq \ker(\alpha^3) \subseteq \dots$ and apply the noetherian condition.
19. (a) To see that θ is multiplicative, let $\theta(r) = [r_{ij}]$ and $\theta(s) = [s_{ij}]$, so that $u_i r = \sum_{j=1}^n r_{ij} u_j$ and $u_i s = \sum_{j=1}^n s_{ij} u_j$ for each i . Compute $u_i r s$.

EXERCISES 11.2 THE WEDDERBURN-ARTIN THEOREM

2. (a) Show that $R = L \oplus M$ for some left ideal, and $1 = e + f$ where $e \in L$ and $f \in M$.
3. (a) The axioms are routinely verified.
5. Each $\frac{M}{N_i}$ is simple.
7. Use the preceding exercise and Lemma 3 §11.1.
9. Show that R is simple as a left R -module.
11. The left ideals of the ring R/A are simultaneously left R -modules and left R/A -modules with the same action.
13. (2) \Rightarrow (1) If $0 \neq x \in Re$ show that $xax \neq 0$, $a \in R$. Show that $Rx = Re$.
15. Use Theorem 1(1) §11.1.
16. (a) $(ML)^2 = MLML$. (b). $(Ar)^2 = ArAr$.
17. Use Exercise 4.
19. Use Lemma 3 §3.3.
20. (a) Use the definition of domain.
21. Use Lemma 9 and Theorem 3.
22. (a) and (c) Use Schur's lemma.

EXERCISES A COMPLEX NUMBERS

1. (a) $x = 3$ (c) $x = 0, x = 4i$
2. (a) $7 - 9i$ (c) $\frac{-6}{13} + \frac{4}{13}i$ (e) $-i$ (g) -4
3. (a) $1 - 3i$ (c) $\pm \frac{1}{\sqrt{2}}(1 - i)$ (e) $2 + 3i$
5. (a) Unit circle (c) Line $y = x$ (e) $\{r \in \mathbb{R} \mid r \geq 0\}$
10. (a) $3\sqrt{2}e^{-\pi i/4}$ (c) $2e^{5\pi i/6}$ (e) $7e^{-\pi i/2}$
11. (a) -3 (c) $-\sqrt{2} + \sqrt{2}i$ (e) $-\frac{1}{\sqrt{2}} - \frac{1}{\sqrt{2}}i$
12. (a) $-2 - 2\sqrt{3}i$ (c) 16 (e) -64
14. (a) $\pm \frac{1}{\sqrt{2}}(1 + i), \pm \frac{1}{\sqrt{2}}(1 - i)$ (c) $3i, \frac{3}{2}(\sqrt{3} - i), \frac{3}{2}(-\sqrt{3} - i)$
19. If $f(x) = z_0 + z_1x + z_2x^2 + \dots + z_nx^n$, the coefficient of x^k in $f(x)\bar{f}(x)$ is $z_0\bar{z}_k + z_1\bar{z}_{k-1} + \dots + z_{k-1}\bar{z}_1 + z_k\bar{z}_0 = (z_0\bar{z}_k + z_k\bar{z}_0) + (z_1\bar{z}_{k-1} + z_{k-1}\bar{z}_1) + \dots + z_{k/2}\bar{z}_{k/2}$, where the last term is real but missing if k is odd. Each of the other summands is also real, being a complex number plus its conjugate.

EXERCISES B MATRIX ARITHMETIC

1. Use A^{-1} .
2. (a) Use the definition of matrix multiplication.
3. Compute.
5. I and $-I$.
7. In general, show $(A + B)(A - B) = A^2 + AB - BA - B^2$.
9. $A^{-1} = A^2$.
11. (a) In general, if $AC = I = CA$ then A is invertible and $A^{-1} = C$.
(c) If $I + BA$ is invertible, compute $(I + AB)(I - A(I + BA)^{-1}B)$.
13. (a) Use Theorem 7.
15. (a) Use the definition of matrix multiplication.
(c) $a_{ij}E_{ij}$ has a_{ij} in the (i, j) -entry and zeros elsewhere.

EXERCISES C ZORN'S LEMMA

2. (a) Let $\mathcal{S} = \{X \subseteq M \mid X \text{ is a submodule and } K \cap X = 0\}$. Then \mathcal{S} is nonempty because $0 \in \mathcal{S}$, so let $\{X_i \mid i \in I\}$ be a chain from \mathcal{S} and put $U = \bigcup_{i \in I} X_i$. It is clear that U is a submodule, and $K \cap U = 0$ because $K \cap U \subseteq K \cap X_i = 0$ for each i . Hence U is an upper bound for the chain $\{X_i \mid i \in I\}$, so \mathcal{S} contains maximal members by Zorn's lemma.

Index

- Abel, N.H. (1802–1829), 69, 202, 377, 397, 413, 432, 435
abelian group, 77
 finite p-groups, 341
 fundamental theorem, 346
 primary decomposition theorem, 339
absolute value, 201
absolute value of a complex number, 473
action of a mapping, 10
additive notation, 71
al-Khowarizmi, M. (c.825), 202
Alexandroff, P.S. (1896–1982), 196
algebra, 448
 regular representation, 457
algebraic closure, 289, 296
algebraic element, 283, 286
algebraic numbers, 289, 296
algebraically closed field, 296
alternating group, 62, 78, 128
alternating polynomial, 245
annihilator, 187, 456
 in a module, 327, 337
 in a ring, 182, 314
annihilator ideal, 182
Archimedes of Syracuse (272–212 BC), 224, 307
Artin, E. (1898–1962), 160, 422, 447, 448, 466
artinian, left, 450
associative law, 70
 for composition, 13
 general, 72
automorphism
 Frobenius, 178
 inner, 104, 105, 140, 168
 of rings, 167
automorphism group, 104
axiom of choice, 489
axiomatic method, 3
axioms, 3
Bézout domain, 272
Bézout, E. (1730–1783), 34, 272
basis
 of a module, 330, 462
 of a vector space, 278, 458, 487
 standard, 278, 331, 463
Bass, H. (1932–), 468
BCH codes, 319
bijection, 11, 54
bijection theorem, 22
bijective mapping, 11

- binary operation, 70
 - associative, 70
 - closed under, 70
 - commutative, 70
 - componentwise, 80
 - unity for, 70
- binomial
 - coefficient, 26
 - theorem, 27, 163
- bits, 144, 310
- Boole, G. (1815–1864), 170
- boolean ring, 170
- Brauer's lemma, 465
- Brauer, R. (1901–1977), 465
- Burnside's lemma, 383
- Burnside, W. (1852–1927), 383, 399, 405
- Burnside-Wielandt theorem, 405
- cancellation, 81
- Cantor, G. (1845–1918), 297
- Cantor, G. (1845–1918) , 284
- Cardano, G. (1501–1576), 412, 435
- cartesian product, 8
- casting out nines, 46
- Cauchy's theorem, 360
- Cauchy, A.L. (1789–1857), 360, 362, 383, 432
- Cauchy-Frobenius lemma, 383
- Cayley table, 71
- Cayley's theorem, 106
 - extended, 365
- Cayley, A. (1821–1895), 69, 71, 106, 202, 362, 447
- center
 - of a group, 87, 123
 - of a ring, 165
- central series, 402, 403
- central series of a group, 404
 - ascending, 402
 - descending, 403
- centralizer, 89, 131, 359
- chain condition
 - ascending, 448
 - descending, 448
 - on principal ideals, 255
- character of a group, 422
 - Dedekind's lemma, 422
 - independent, 422
- characteristic, 193
- characteristic of a ring, 163
- characteristic subgroup, 124, 130
- Chevalley, C. (1909–1984), 467
- Chinese remainder theorem, 48, 98, 195
- circle group, 77
- class equation, 359
- code
 - (n,k)-code, 148
 - BCH, 319
 - binary, 143
 - check bits, 145
 - coset leader for, 149
 - cyclic, 311, 312
 - error correction, 147, 148
 - error detection, 147, 148, 317
 - Hamming (7,4), 151
 - idempotent generator, 322
 - matrix description, 151
 - matrix generator, 315
 - maximum likelihood decoding, 144
 - minimal generator, 312
 - nearest neighbor, 146
 - orthogonality theorem, 155
 - parity check, 145
 - perfect, 148
 - polynomial/word form, 311
 - standard array for, 150
 - standard generator matrix, 152
- syndrome, 154
- syndrome decoding, 155
- words, 144
- codomain of a mapping, 10
- Cohen, P. (1934–), 489
- combinatorics application, 382
- commutative law, 70
- commutative ring, 160
 - simple, 184
- commutator, 134
 - subgroup, 134
- complemented module, 459
- completing the square, 47
- complex number, 6, 471
 - absolute value/modulus of, 473
 - conjugate of, 473
 - DeMoivre's theorem, 476
 - distance formula, 477
 - exponential form, 474
 - inverse of, 473
 - modulus, 473
 - multiplication rule, 475
 - operations, 472
 - polar form of, 475
 - real/imaginary part, 472

- roots of unity, 476
- triangle inequality, 477
- complex plane, 471
 - imaginary axis, 472
 - real axis, 472
- component of an n -tuple, 8
- componentwise operation, 80, 160
- composite of mappings, 12
- composition factors, 390
- composition length, 390
- composition series
 - group, 389
 - module, 451
- conclusion, 2
- congruence modulo n , 43, 48, 50
- conjugacy class, 358
- conjugate elements, 358
- conjugate in quadratic integers, 268
- conjugate of a complex number, 473
- conjugate subgroups, 88, 105, 123, 358
- constructible number, 305
- containment
 - proper, 5
- contradiction, 2
- contrapositive, 5
- converse, 3
- convolution, 249
- core of a subgroup, 131, 364
- corner ring, 165, 454
- correspondence theorem, 353
- coset
 - double, 117
 - left/right, 109
 - map, 132
- coset decoding, 149
- coset leader, 149
- coset map, 132
- counterexample, 3
- cubic formula, 435
- cubic polynomial, 205
- cycle, 58
 - decomposition theorem, 60, 62
 - length, 58
 - structure, 60
- cycle index, 385
- cyclic codes, 311, 312
- cyclic group, 82, 91, 94
 - fundamental theorem, 95
 - of order n , 82
- cyclic subgroup, 91
 - generator of, 91
- cyclotomic polynomial, 221, 442, 445
- D'Alembert, J. LeR. (1717–1783), 324
- De Morgan, A. (1806–1871), 24, 412
- Dedekind's lemma, 422
- Dedekind, R. (1831–1916), 159, 181, 185, 196, 252, 271, 297, 422
- Dedekind-Artin theorem, 424
- degree of a polynomial, 205
- degree over F , 285, 414
- DeMoivre's theorem, 476
- DeMoivre, A. (1667–1754), 476
- density theorem, 467
- derivative, formal, 299
- derived
 - series, 396
 - subgroup, 134
 - higher, 396
- Descartes, R. (1596–1650), 8, 202
- Dickson, L. E. (1874–1954), 398
- dicyclic group, 375
- difference of two sets, 7
- dihedral group, 113, 121
- dimension of a vector space, 279
- Dirac, P.A.M. (1902–1984), 349
- direct product, 80
- direct sum
 - characterization, 328
 - external, 325
 - internal, 329
- Dirichlet, P.G.L. (1805–1859), 40
- discriminant
 - of a cubic, 435
 - of a polynomial, 442
 - of a quadratic, 217, 434
- disjoint sets, 18
- divisible group, 335
- division algorithm, 266
 - for integers, 32
 - for polynomials, 207
- division ring, 166, 444
- divisor, 33, 221, 252
 - common, 33
 - greatest common, 33, 221
- domain, 172
 - Ore, 177
- domain of a mapping, 10

- Doyle, A.C. (1859–1930), 67
 duplicating a cube, 304, 307
- Eisenstein criterion, 220
 element of a set, 5
 elementary divisors, 346
 elementary symmetric polynomial, 242, 309, 439
 elements of sets, 5
 embedding theorem, 177
 empty set, 6
 endomorphism, 452
 endomorphism ring, 452
 equivalence
 afforded by a partition, 19
 class, 17
 kernel of, 18
 logical, 3
 natural mapping, 19
 quotient set of, 19
 relation, 17
 Euclid of Alexandria (ca 330–275 BC), 35, 37, 40
 euclidean algorithm, 35
 euclidean domain, 270
 Euler function, 114
 Euler, L. (1707–1783), 40, 114, 202, 362, 435
 evaluation map, 208
 exponent laws, 72, 74, 81
 extension of fields, 283, 291
 abelian, 434
 algebraic, 283, 289
 closures in, 426
 cyclic, 434
 dimension of, 283
 F-automorphism, 413
 finite, 283, 287, 288
 Galois, 427, 441
 Galois group, 413
 intermediate field, 425
 main theorem of Galois theory, 429
 norm, 434
 normal, 298
 primitive element theorem, 419
 radical, 435
 separable, 418, 419
 separable closure, 422
 simple, 284, 302, 419
 trace, 434
- factor group, 131, 132
 factor modules, 327
 factor ring, 181
 ideals in, 183
 factor theorem, 209
 factorial, 26
 factorization, 252
 trivial, 252
 FC-group, 363
 Feit, W. (1930–2004), 129, 399
 Fermat's theorem, 50
 Fermat, P. (1601–1665), 1, 50, 51, 115
 Ferrari, L. (1522–1565), 412, 435
 Ferro, S. (1522–1565), 412
 field, 49, 166
 algebraic closure, 289
 algebraic element, 283
 algebraically closed, 296
 Dedekind's lemma, 422
 Dedekind-Artin theorem, 424
 extension, 283
 finite, 298
 Galois, 300
 minimal polynomial of an element, 285, 414
 of algebraic numbers, 289
 of constructible numbers, 305
 of quotients, 177
 perfect, 421
 prime, 194
 splitting, 292, 418
 transcendental element, 283
 field of quotients, 177
 figure, 117
 motion of, 117
 finite field, 298
 cyclic unit group, 302
 existence, 300
 primitive elements, 302
 subfields, 301
 uniqueness, 300
 finite subgroup test, 87
 finitely generated group, 96
 module, 330
 Fitting's lemma, 457
 Fitting, H. (1906–1938), 408, 457
 Fitting's theorem, 408
 fixed field of a group, 424
 four square identity, 175
 Fourier, J.B.J. (1768–1830), 432
 Frattini argument, 373
 Frattini subgroup, 406

- Frattini, G. (1852–1925), 373, 406
 free module, 331, 462
 Frobenius
 automorphism, 178, 191, 299, 300
 endomorphism, 191
 theorem, 371
 Frobenius, G. (1849–1917), 371, 372, 377, 383, 399, 447
 fully invariant submodule, 461
 fundamental identities of a mapping, 14
 fundamental theorem
 finitely generated modules over a PID, 345
 for finite abelian groups, 346
 for vector spaces, 278
 of algebra, 309, 471
 of cyclic groups, 95
 of Galois theory, 429
 of symmetric polynomials, 242, 309
- Gödel, K. (1906–1978), 489
 Galois connection, 426
 Galois extension, 427, 441
 Galois field, 300
 Galois group of a polynomial, 436
 Galois group of an extension, 413
 Galois theory
 main theorem, 429
 Galois' criterion, 436
 Galois, E. (1811–1832), 69, 202, 298, 413, 432, 435, 436, 438
 Gauss, C.F. (1777–1855), 164, 216, 218, 224, 251, 252, 261, 308, 362
 formula, 25, 32
 Gaussian integers, 164
 gcd, 33, 34, 36, 39, 258
 integers, 36
 polynomials, 222
 general associativity, 72
 general linear group, 80
 generator of a cyclic group, 82, 91
 generators
 of a field extension, 284
 of a group, 78, 100
 of a subgroup, 96
 Goldbach, C. (1690–1764), 40
 Goldie, A.W. (1920–2005), 468
 Grassmann, H.G. (1809–1877), 159, 447
 greatest common divisor, 33, 39, 258
 group, 76
 abelian, 77
 actions (G-sets), 365
 alternating, 62, 78, 128
 automorphism, 104
 Burnside–Wielandt theorem, 405
 Cauchy's theorem, 360
 Cauchy–Frobenius Lemma, 383
 Cayley's theorem, 106
 center of, 87
 central series, 404
 character, 422
 circle group, 77
 class equation, 359
 composition series, length, factors, 389
 core of a subgroup, 131, 364
 correspondence theorem, 353
 cosets of a subgroup, 109
 cyclic, 82, 94
 dicyclic, 375
 dihedral, 113, 121
 direct product, 80
 divisible, 335
 extended Cayley theorem, 365
 extension problem, 388, 391
 factor group of, 132
 FC-group, 363
 finite abelian p-groups, 341
 finite p-groups, 338
 finitely generated, 96
 Frattini subgroup, 406
 G-set, 365
 general linear, 80
 Hall's theorem, 399
 holomorph, 357
 homomorphism, 99
 image, 352
 isomorphic, 82
 isomorphism, 102
 isomorphism theorem, 138
 Jordan–Hölder theorem, 390
 Klein, 83
 lattice diagram, 87
 maximal normal subgroup, 355
 metabelian, 357
 metacyclic, 355
 nilpotent, 404
 nongenerators of, 407
 normalizer, 359
 octic, 113
 of motions, 79, 117
 of units, 79, 165
 p-group, 360
 permutation, 79
 polycyclic, 401
 Prüfer, 449

- group (*Continued*)
 product of subgroups, 125
 product of subsets, 350
 projective special linear, 398
 quaternion, 127
 relations in, 78
 Schur's theorem, 381
 Schur-Zassenhaus theorem, 382
 second isomorphism theorem, 350
 semidirect product, 379
 simple, 128, 398
 simple factor, 355
 solvable, 395, 397, 436
 special linear, 86, 138
 subgroup of, 86
 Sylow theorems, 371
 symmetric, 54, 56, 78, 94
 third isomorphism theorem, 355
 torsion, 136
 torsion-free, 136
 translations of, 357
 group action, 365
 by conjugation, 366
 by multiplication, 366
 faithful, 370
 fixed element, 367
 fixed subset, 368
 fixer of, 367
 G-morphisms, 371
 orbit decomposition theorem, 368
 orbit in, 367
 stabilizer, 368
 transitive, 370
 trivial, 366
 group of units, 165
 Hölder, O. (1859–1937), 390
 Hall's theorem, 399
 Hall, P. (1904–1982), 399
 Hall, P. (1904–1982), 399
 Hamilton, W.R. (1805–1865), 115, 159, 173, 447
 Hamming, R. (1915–1998), 143, 146
 (7,4)-code, 151
 bound, 148
 code, 156
 distance, 146
 weight, 146
 Hankel, H. (1839–1873), 412
 Hardy, G.H. (1877–1947), 364
 Hermite, C. (1822–1901), 283
 Hertz, H.R. (1857–1894), 202
 higher derived subgroups, 396
 Hilbert, D. (1862–1943), 5, 159, 196
 Hobbes, T. (1588–1679), 304
 homomorphism, 100
 fixed element of, 294
 general ring, 189
 group, 99
 image of, 100, 137, 192, 326
 kernel of, 137, 192, 326
 module, 326
 preimage of, 141
 ring, 189
 trivial, 99, 138
 Hopkins-Levitzky theorem, 450
 hypothesis, 2
 ideal, 181
 annihilator, 187, 314
 left, 187, 488
 left/right, 326
 maximal, 184
 maximal left, 488
 prime, 182
 principal, 181
 proper, 181
 zero, 181
 idempotent, 75, 165, 453
 lifting, 468
 identity
 homomorphism, 326
 mapping, 13
 permutation, 55
 image of a homomorphism, 100, 137, 192, 326
 image of a mapping, 12
 implication, 2
 inclusion mapping, 137
 indeterminant, 203
 index of a subgroup, 111
 induction
 definition by, 29
 hypothesis, 24
 mathematical, 24
 principal of, 24
 strong, 28
 inductive definition, 29
 inductive set, 457
 inner automorphism, 105, 140, 168
 integers, 5
 prime, 36
 relatively prime, 36
 integers modulo n, 44, 45, 47, 49

- integral domain, 172
- ACCP, 255
- as factor ring, 183
- associates in, 253
- Bézout, 272
- embedding theorem, 177
- euclidean, 270
- field of quotients, 177
- greatest common divisor (gcd), 258
- irreducible in, 254
- least common multiple (lcm), 258
- ordered, 199
- positive elements, 199
- prime element of, 257
- prime ideal of, 488
- principal ideal domain (PID), 264
- reducible in, 254
- unique factorization domain (UFD), 256
- well-ordered, 200
- intermediate field of an extension, 425
- internal direct sum, 329
- intersection of sets, 7
- invariant basis number (IBN), 332
- invariant factors, 346
- inverse
 - in a monoid, 73
 - left/right, 76
 - of a complex number, 473
 - permutation, 56
- inverse of a mapping, 14
- invertibility theorem, 15
- irreducible, 215
- irreducible, 254
 - polynomial, 215
- isometry, 119
- isomorphic groups, 82, 102
- isomorphic rings, 167
- isomorphism, 102
 - of rings, 167
- isomorphism theorem
 - group, 138
 - module, 327
 - ring, 192
 - second group, 350
 - second ring, 198
 - third group, 355
 - third ring, 198
- Jacobson radical, 467
- Jacobson's theorem, 444
- Jacobson, N. (1910–1999), 444, 467
- Jordan, C. (1832–1922), 398
- Jordan, C. (1838–1922), 390
- Jordan–Hölder theorem
 - group, 390
- Jordan–Hölder theorem, 461
 - module, 451
- Kaplansky, I. (1917–2006), 426
- kernel of a homomorphism, 137, 192, 326
- Klein group, 83
- Klein, F. (1849–1925), 83
- Kronecker delta, 486
- Kronecker's theorem, 233, 291
- Kronecker, L. (1823–1891), 23, 233, 271, 291, 296, 346
- Kummer, E.E. (1810–1893), 181, 251, 271, 297
- Lagrange
 - four square identity, 175
 - interpolation, 214
 - polynomials, 214
- Lagrange's theorem, 109, 111, 114, 115
- Lagrange, J.L. (1736–1813), 108, 111, 115, 175, 202, 214, 413, 432, 435
- Laplace, P.S. de (1749–1827), 483
- lattice diagram, 87
- lcm, 39, 258
- leading coefficient, 205
- least common multiple, 39, 258
- left ideal, 326
- Legendre, A.-M. (1752–1833), 432
- Lindemann, F. (1852–1939), 283, 307
- linear combination, 33, 277
 - trivial, 277
- linear polynomial, 205
- Lobachevski, N.I. (1793–1865), 310
- localization, 188
- logically equivalent, 3
- Möbius function, 446
- Möbius inversion formula, 446
- main theorem of Galois theory, 429
- Mal'cev, A.I. (1909–1967), 177
- mapping, 10, 15
 - action of, 10
 - bijective, 11
 - composite, 12
 - constant, 16
 - domain, codomain, 10
 - fundamental identities, 14
 - identity, 13
 - image, 10

- mapping (*Continued*)
 image of, 12
 inverse of, 14
 natural, 19
 one-to-one, 11
 onto, 11
 structure preserving, 99, 189, 190
 surjective, 11
 well-defined, 10
- matrix, 161, 162, 479
 identity, 162, 481
 m by n, 479
 main diagonal, 162, 481
 operations, 162
 parity check, 154
 product, 162, 480
 ring, 162
 similarity, 168
 square, 479
 zero, 162, 479
- matrix rings, 162
 ideals in, 185
- matrix units, 184
- Maurolico, Francesco (1494–1575), 24
- maximal
 subgroup, 405
 maximal ideal, 184
 maximal normal subgroup, 355
- McKay, J.H., 369
- metacyclic group, 355
- minimal polynomial over F, 285, 414
- modular irreducibility test, 220
- modular law
 for subgroups, 350
- module, 324, 325
 annihilator in, 327
 artinian, 448
 basis of, 330, 462
 complemented, 459
 composition series, length, factors, 451
 direct sum, external, 325
 direct sum, internal, 329
 endomorphism ring, 452
 factor modules, 327
 finite dimensional, 456
 finitely generated, 330
 free, 331, 462
 generating set, 330, 462
 homomorphism, 326
 indecomposable, 456
 independent set, 462
 independent subset, 330
- invariant basis number (IBN), 332
- isomorphic, 326
 isomorphism, 326
 isomorphism theorem, 327
 left, 325
 morphism, 326
 noetherian, 448
 over a PID, 335
 principal, 326
 projective, 332, 464
 rank, 333
 rank theorem, 332
 semisimple, 448, 458, 460, 462
 simple, 334, 450
 submodule, 326
 sum, 328
 torsion-free, 327
 trivial morphism, 326
- modules over a PID, 335
 annihilators in, 337
 decomposition of p-modules, 339
 free if and only if torsion-free, 335
 fundamental theorem, 345
 order of elements, 336
 p-modules, 339
 p-primary component, 337
 primary decomposition theorem, 337
 submodule theorem, 343
 torsion submodule, 336
- modulus, 43
 modulus of a complex number, 473
- monic polynomial, 205, 215, 219, 221–223
- monoid, 70
 commutative, 70
- monomial, 240
- morphism
 identity, 326
 image of, 326
 kernel of, 326
 module, 326
- motion of a figure, 117
- multiplication rule for complex numbers, 475
- multiplication theorem, 287
- multiplicative notation, 71
- natural mapping, 19
- natural numbers, 5
- negative, 45, 73, 160
- Newton identities, 244
- Newton, I. (1642–1727), 224
- nil radical, 188

- nilpotency class, 404
- nilpotent, 166, 488
 - ideal, 465
- nilpotent group, 404
 - Burnside–Wielandt theorem, 405
- Noether, E. (1882–1935), 159, 196, 448
- noetherian rings, 196
 - left, 450
- nongenerator in a group, 407
- norm in quadratic integers, 268
- normal closure, 131
- normal subgroup, 122, 131
 - test for, 123
- normalizer, 359
- normalizer of a subgroup, 130
- number
 - of elements in a set, 12
- octic group, 113
- one-to-one mapping, 11
- onto mapping, 11
- orbit, 367
- orbit decomposition theorem, 367, 368
- order
 - of a group, 77
- order of an element
 - finite, 92
 - in a module over a PID, 336
 - infinite, 92
- ordered integral domain, 199
- ordered n-tuple, 8
 - component of, 8
- ordered pair, 7
- Ore domains, 177
- Ore, O. (1899–1968), 177
- Oresme, Nicole (1323–1382), 8
- orthogonality theorem, 155
- p-group, 360
 - finite, 339
 - finite abelian, 338
- p-module, 339
 - direct sum decomposition, 339
 - elementary divisors, 340
 - type, 341
 - uniqueness of decomposition, 339
- parity, 61
- parity check matrix, 154, 315
- parity theorem, 61, 63
- partial fraction expansion, 237
- partial order, 486
- partially ordered set (poset), 486
 - inductive, 487
 - upper bound in, 487
- partition, 18
 - cells of, 18
 - singleton, 19
 - theorem, 19
 - trivial, 19
- Pascal, B. (1623–1662), 27, 30
 - identity, 26
 - triangle, 27
- Peano axioms, 29
- Peano, Giuseppe. (1858–1932), 29
- Pell's equation, 269
- permutation, 54
 - cycle, 58
 - cycle index, 385
 - disjoint, 57, 58, 60
 - even, 61
 - fixed/moved element of, 57
 - identity, 55
 - inverse, 56
 - odd, 61
 - of a set, 79
 - sign of, 138
- permutation group, 79
- PID, 264, 335, 343
 - and UFD's, 265
 - module decomposition, 335
 - module is free iff torsion-free, 335
 - primes in, 266
- PID (principal ideal domain), 264, 335
- pigeonhole principle, 4
- Poincaré, H. (1854–1912), 99, 117, 447
- pointwise operations, 161
- Poisson, S.-D. (1781–1840), 432
- polycyclic group, 401
- polynomial, 203, 240
 - alternating, 245
 - coefficients, 203
 - constant, 204
 - constant coefficient, 204
 - cubic, 205
 - cyclotomic, 221, 442, 445
 - degree, 205, 240
 - derivative of, 299
 - Eisenstein criterion, 220
 - elementary symmetric, 242
 - equality, 204
 - error-locator, 321
 - evaluation theorem, 208
 - even/odd, 282

- polynomial (*Continued*)
 factor is a field, 232
 factor rings, 230
 factor theorem, 209
 fundamental theorem of symmetric polynomials, 242, 309
 gcd, 222
 homogeneous, 240
 homogeneous components, 240
 irreducible, 215
 Kronecker's theorem, 233
 Lagrange, 214
 leading coefficient, 205
 least common multiple, 235
 lexicographic order, 241
 linear, 205
 minimal, 285, 414
 modular irreducibility test, 220
 monic, 205, 215, 219, 221–223
 negative of, 204
 Newton identities, 244
 partial fraction expansion, 237
 primitive, 260
 principle ideal domain, 227
 proper factorization, 218
 quadratic, 205
 quartic, 205
 quintic, 205
 rational forms, 237, 440
 rational roots theorem, 211
 reducible, 215
 relatively prime, 222, 235
 remainder theorem, 209
 repeated root, 300
 ring, 203
 roots of, 210
 separable, 418
 several variables, 239
 solvable, 435
 splits, 292
 symmetric, 241, 309, 439
 symmetric rational forms, 440
 unique factorization theorem, 223
 zero, 204
- positive elements, 199
 positive numbers, 6
 power of an element, 72
 Prüfer group, 449
 preimage, 141
 primary component, 337, 372
- primary decomposition theorem
 for finite abelian groups, 339
 for modules over a PID, 337
- prime, 32, 36, 37, 40
 in a PID, 337
- prime factorization, 36
 theorem, 37
- prime fields, 194
- prime ideal, 182
- prime number, 3
- prime power, 38
- prime ring, 469
- primitive element, 302
- primitive element theorem, 288, 419
- primitive polynomial, 260
- primitive root modulo p, 302
- primitive roots of unity, 302, 437, 476
- principal ideal, 181
- projection, 107, 334
- projections, 332
- projective module, 332, 464
- projective special linear group, 398
- proof, 2
 by contradiction, 2
 direct method, 2
 reduction to cases, 2
- proper
 ideal, 181
 subgroup, 86
 submodule, 459
- quadratic
 integers, 266, 267
 quadratic formula, 202, 217, 397, 434
 quadratic polynomial, 205
 quartic polynomial, 205
 quaternion group, 127
 quaternions, 174, 179
 conjugate, 174
 norm, 174
 quintic polynomial, 205
 quotient, 32
 quotient set, 19
- radian measure of an angle, 474
 radical extension, 435
 rank of a module, 333
 rank theorem for modules, 332
 rational expression, 190
 rational forms, 237, 440
 rational numbers, 5
 rational roots theorem, 211

- real numbers, 6
- recursion theorem, 29, 490
- reducible, 254
 - in an integral domain, 254
- reducible polynomial, 215
- regular n-gon, 120
- regular representation, 457
- relation, 17
- relatively prime, 36, 37
 - integers, 36
 - polynomials, 222
- remainder, 32
- remainder theorem, 209
- residue class, 43
- residue modulo n, 43
- ring, 160
 - automorphism, 167
 - binomial theorem, 163
 - boolean, 170
 - center, 165
 - characteristic of, 163
 - Chinese remainder theorem, 195
 - commutative, 160
 - corner ring, 165, 454
 - decomposition theorem, 195
 - density theorem, 467
 - direct product, 161
 - division, 166, 444
 - endomorphism, 452
 - factor, 181
 - general, 160, 194
 - group of units, 165
 - homomorphism, 189
 - ideal of, 181
 - idempotent in, 165, 453
 - isomorphism, 167
 - isomorphism theorem, 192
 - Jacobson radical, 467
 - left artinian, 450
 - left noetherian, 450
 - lifting idempotents, 468
 - local, 188
 - maximal left ideal, 488
 - negative in, 160
 - nil radical of, 188, 488
 - nilpotent ideal, 465
 - nilpotent in, 166, 488
 - noetherian, 196
 - of functions, 161
 - of matrices, 161
 - opposite, 169
 - polynomial, 203
 - prime, 469
 - prime ideal of, 488
 - semiperfect, 468
 - semisimple, 467
 - simple, 183–185
 - subring, 164
 - subtraction, 163
 - unit, 165
 - unity of, 160
 - upper triangular, 164
 - zero ring, 161
 - root
 - multiplicity, 210
 - of a polynomial, 210
 - of unity, 77, 302, 437, 476
 - rational, 211
 - repeated, 300
 - Ruffini, P. (1765–1822), 397, 413
 - Russell, B. (1872–1970), 9, 159, 489
 - Schur's lemma, 334, 454
 - Schur's theorem, 381
 - Schur, I. (1875–1941), 381, 454
 - Schur-Zassenhaus theorem, 382
 - second isomorphism theorem
 - for groups, 350
 - for rings, 198
 - semidirect product, 379, 380
 - semiperfect ring, 468
 - semisimple module, 448, 458, 460, 462
 - homogeneous, 462
 - homogeneous component, 461
 - semisimple rings, 467
 - separable closure, 422
 - separable extension, 418
 - sequence, 29, 203, 248, 490
 - recursively defined, 490
 - set, 5
 - cartesian product, 8
 - containment, 5
 - difference, 7
 - disjoint, 18
 - element of, 5
 - empty, 6
 - equality, 5
 - infinite, 6
 - intersection, 7
 - operations on, 7
 - proper containment, 5
 - singleton, 6
 - subset of, 5
 - union, 7

- Shannon, C.E. (1916–2001), 143, 394
 sign of a permutation, 66, 138
 similar matrices, 168
 simple
 group, 128, 398
 module, 334, 450
 ring, 183–185
 solvable group, series, 395
 solvable polynomial, 435, 436
 span of vectors, 277
 special linear group, 86, 138
 projective, 398
 Split mapping, 463
 splitting field, 292, 418
 existence, 292
 uniqueness, 294
 square free, 38
 stabilizer, 368
 Steinitz exchange lemma, 279
 subfield, 283
 subgroup, 86
 characteristic, 124, 130
 conjugate, 88, 123, 358
 cyclic, 91
 derived, 134, 396
 Frattini, 406
 generated by a set, 96
 generators, 89
 index of, 111
 maximal, 405
 maximal normal, 355
 normal, 88, 122, 131
 proper, 86
 self-conjugate, 88
 subnormal, 410
 test, 86
 torsion, 136
 trivial, 86
 unconnected set of, 352
 subgroup test, 86
 submodule, 326
 direct sum, 328
 fully invariant, 461
 maximal, 459
 principal, 326
 proper, 459
 sum of, 328
 submodule theorem, 343
 subset, 5
 subspace test, 276
 surjective mapping, 11
 Sylow p-subgroup, 372
 number of, 374
 Sylow theorems, 371
 Sylow's first theorem, 372
 Sylow's second theorem, 373
 Sylow's third theorem, 374
 Sylow, L. (1832–1918), 371, 377
 symmetric group, 54, 56, 78, 94
 symmetric polynomial, 241, 309, 439
 elementary, 309, 439
 symmetric rational forms, 440
 symmetry of a figure, 119
 syndrome, 154
 Tartaglia, N. (1500–1557), 412, 435
 theorems, 3
 third isomorphism theorem
 for groups, 355
 for rings, 198
 Thompson, J.G. (1932–), 129, 399
 torsion group, 136
 torsion subgroup, 136
 torsion submodule, 334, 336
 torsion-free
 element, 327
 module, 327, 334
 torsion-free group, 136
 transcendental element, 283
 transposition, 60, 61
 trisecting an angle, 304, 307
 trivial factorization, 215, 252
 trivial homomorphism, 99, 138, 326
 trivial linear combination, 277
 trivial subgroup, 86
 Tucker, A., 143
 UFD, 256, 265
 characterization, 259
 Gauss' lemma, 261
 polynomials, 261
 unconnected subgroups, 352
 union of sets, 7
 unique factorization theorem, 223
 unit, 73, 165
 circle, 474
 in a monoid, 79
 properties, 74
 unit circle, 474
 unity, 70
 left/right, 75
 unity for a binary operation, 70
 upper triangular, 164

- variety of groups, 401
- vector space, 275
 - basis of, 278, 458, 487
 - dependence in, 277
 - dimension, 279
 - dimension theorem, 282
 - finite dimensional, 277
 - fundamental theorem, 278
 - invariance theorem, 279
 - linear combination, 277
 - linear independence, 277, 458, 487
 - linear transformation, 282
 - scalar multiples, 275
 - spanning set, 277, 458, 487
 - subspace, 276
 - zero space, 276
- Venn diagrams, 7
- Venn, J. (1834–1883), 7
- Voltaire–F.M.A. (1694–1778), 274
- Wedderburn's theorem, 444
 - on division rings, 173
 - on simple rings, 455
- Wedderburn, J.H.M. (1882–1948), 159, 173, 442, 443, 447, 455, 466
- Wedderburn-Artin Theorem, 457
- Weierstrass, K. (1815–1897), 297
- Weil, A. (1906–1998), 388
- well-defined mapping, 10, 15, 131
- well-ordered integral domain, 200
- well-ordering
 - axiom, 28
- Weyl, H. (1885–1955), 196, 432, 455
- Whitehead, A.N. (1861–1947), 1, 159
- Wielandt, H. (1910–2001), 377, 405
- Wiles, A., 51
- Wilson's theorem, 49, 97
- Witt, E., (1911–1991), 443
- word
 - empty, 75
 - justaposition, 75
 - length, 75
- Zassenhaus, H. (1912–1991), 381, 394
- Zermelo, E. (1871–1953), 489
- zero
 - ideal, 181
 - matrix, 162
 - vector space, 276
- Zorn's lemma, 333, 457, 487
- Zorn, M. (1906–1993), 487