MODERN ALGEBRA
WITH APPLICATIONS

PURE AND APPLIED MATHEMATICS

A Wiley-Interscience Series of Texts, Monograph, and Tracts

Founded by RICHARD COURANT
Editors: MYRON B. ALLEN III, DAVID A. COX, PETER LAX
Editors Emeriti: PETER HILTON, HARRY HOCHSTADT, JOHN TOLAND

A complete list of the titles in this series appears at the end of this volume.

MODERN ALGEBRA WITH APPLICATIONS

Second Edition

WILLIAM J. GILBERT
University of Waterloo
Department of Pure Mathematics
Waterloo, Ontario, Canada

W. KEITH NICHOLSON
University of Calgary
Department of Mathematics and Statistics
Calgary, Alberta, Canada



A JOHN WILEY & SONS, INC., PUBLICATION

Cover: Still image from the applet KaleidoHedron, Copyright © 2000 by Greg Egan, from his website http://www.netspace.net.au/∼gregegan/. The pattern has the symmetry of the icosahedral group.

Copyright © 2004 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 978-750-4470, 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, e-mail: permreq@wiley.com.

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 within 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:

Gilbert, William J., 1941–
Modern algebra with applications / William J. Gilbert, W. Keith Nicholson.—2nd ed.
p. cm.—(Pure and applied mathematics)
Includes bibliographical references and index.
ISBN 0-471-41451-4 (cloth)
1. Algebra, Abstract. I. Nicholson, W. Keith. II. Title. III. Pure and applied mathematics (John Wiley & Sons : Unnumbered)

QA162.G53 2003
512—dc21

2003049734

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

Preface to the First Edition **ix**

Preface to the Second Edition **xiii**

List of Symbols **xv**

1 Introduction **1**

Classical Algebra, 1
Modern Algebra, 2
Binary Operations, 2
Algebraic Structures, 4
Extending Number Systems, 5

2 Boolean Algebras **7**

Algebra of Sets, 7
Number of Elements in a Set, 11
Boolean Algebras, 13
Propositional Logic, 16
Switching Circuits, 19
Divisors, 21
Posets and Lattices, 23
Normal Forms and Simplification of Circuits, 26
Transistor Gates, 36
Representation Theorem, 39
Exercises, 41

3 Groups **47**

Groups and Symmetries, 48
Subgroups, 54

Cyclic Groups and Dihedral Groups, 56
Morphisms, 60
Permutation Groups, 63
Even and Odd Permutations, 67
Cayley's Representation Theorem, 71
Exercises, 71

4 Quotient Groups 76

Equivalence Relations, 76
Cosets and Lagrange's Theorem, 78
Normal Subgroups and Quotient Groups, 82
Morphism Theorem, 86
Direct Products, 91
Groups of Low Order, 94
Action of a Group on a Set, 96
Exercises, 99

5 Symmetry Groups in Three Dimensions 104

Translations and the Euclidean Group, 104
Matrix Groups, 107
Finite Groups in Two Dimensions, 109
Proper Rotations of Regular Solids, 111
Finite Rotation Groups in Three Dimensions, 116
Crystallographic Groups, 120
Exercises, 121

6 Pólya–Burnside Method of Enumeration 124

Burnside's Theorem, 124
Necklace Problems, 126
Coloring Polyhedra, 128
Counting Switching Circuits, 130
Exercises, 134

7 Monoids and Machines 137

Monoids and Semigroups, 137
Finite-State Machines, 142
Quotient Monoids and the Monoid of a Machine, 144
Exercises, 149

8 Rings and Fields 155

Rings, 155
Integral Domains and Fields, 159
Subrings and Morphisms of Rings, 161

New Rings from Old, 164
Field of Fractions, 170
Convolution Fractions, 172
Exercises, 176

9 Polynomial and Euclidean Rings 180

Euclidean Rings, 180
Euclidean Algorithm, 184
Unique Factorization, 187
Factoring Real and Complex Polynomials, 190
Factoring Rational and Integral Polynomials, 192
Factoring Polynomials over Finite Fields, 195
Linear Congruences and the Chinese Remainder Theorem, 197
Exercises, 201

10 Quotient Rings 204

Ideals and Quotient Rings, 204
Computations in Quotient Rings, 207
Morphism Theorem, 209
Quotient Polynomial Rings That Are Fields, 210
Exercises, 214

11 Field Extensions 218

Field Extensions, 218
Algebraic Numbers, 221
Galois Fields, 225
Primitive Elements, 228
Exercises, 232

12 Latin Squares 236

Latin Squares, 236
Orthogonal Latin Squares, 238
Finite Geometries, 242
Magic Squares, 245
Exercises, 249

13 Geometrical Constructions 251

Constructible Numbers, 251
Duplicating a Cube, 256
Trisecting an Angle, 257
Squaring the Circle, 259
Constructing Regular Polygons, 259

Nonconstructible Number of Degree 4, 260
Exercises, 262

14 Error-Correcting Codes 264

The Coding Problem, 266
Simple Codes, 267
Polynomial Representation, 270
Matrix Representation, 276
Error Correcting and Decoding, 280
BCH Codes, 284
Exercises, 288

Appendix 1: Proofs 293

Appendix 2: Integers 296

Bibliography and References 306

Answers to Odd-Numbered Exercises 309

Index 323

PREFACE TO THE FIRST EDITION

Until recently the applications of modern algebra were mainly confined to other branches of mathematics. However, the importance of modern algebra and discrete structures to many areas of science and technology is now growing rapidly. It is being used extensively in computing science, physics, chemistry, and data communication as well as in new areas of mathematics such as combinatorics. We believe that the fundamentals of these applications can now be taught at the junior level. This book therefore constitutes a one-year course in modern algebra for those students who have been exposed to some linear algebra. It contains the essentials of a first course in modern algebra together with a wide variety of applications.

Modern algebra is usually taught from the point of view of its intrinsic interest, and students are told that applications will appear in later courses. Many students lose interest when they do not see the relevance of the subject and often become skeptical of the perennial explanation that the material will be used later. However, we believe that by providing interesting and nontrivial applications as we proceed, the student will better appreciate and understand the subject.

We cover all the group, ring, and field theory that is usually contained in a standard modern algebra course; the exact sections containing this material are indicated in the table of contents. We stop short of the Sylow theorems and Galois theory. These topics could only be touched on in a first course, and we feel that more time should be spent on them if they are to be appreciated.

In Chapter 2 we discuss boolean algebras and their application to switching circuits. These provide a good example of algebraic structures whose elements are nonnumerical. However, many instructors may prefer to postpone or omit this chapter and start with the group theory in Chapters 3 and 4. Groups are viewed as describing symmetries in nature and in mathematics. In keeping with this view, the rotation groups of the regular solids are investigated in Chapter 5. This material provides a good starting point for students interested in applying group theory to physics and chemistry. Chapter 6 introduces the Pólya–Burnside method of enumerating equivalence classes of sets of symmetries and provides a very practical application of group theory to combinatorics. Monoids are becoming more

important algebraic structures today; these are discussed in Chapter 7 and are applied to finite-state machines.

The ring and field theory is covered in Chapters 8–11. This theory is motivated by the desire to extend the familiar number systems to obtain the Galois fields and to discover the structure of various subfields of the real and complex numbers. Groups are used in Chapter 12 to construct latin squares, whereas Galois fields are used to construct orthogonal latin squares. These can be used to design statistical experiments. We also indicate the close relationship between orthogonal latin squares and finite geometries. In Chapter 13 field extensions are used to show that some famous geometrical constructions, such as the trisection of an angle and the squaring of the circle, are impossible to perform using only a straightedge and compass. Finally, Chapter 14 gives an introduction to coding theory using polynomial and matrix techniques.

We do not give exhaustive treatments of any of the applications. We only go so far as to give the flavor without becoming too involved in technical complications.

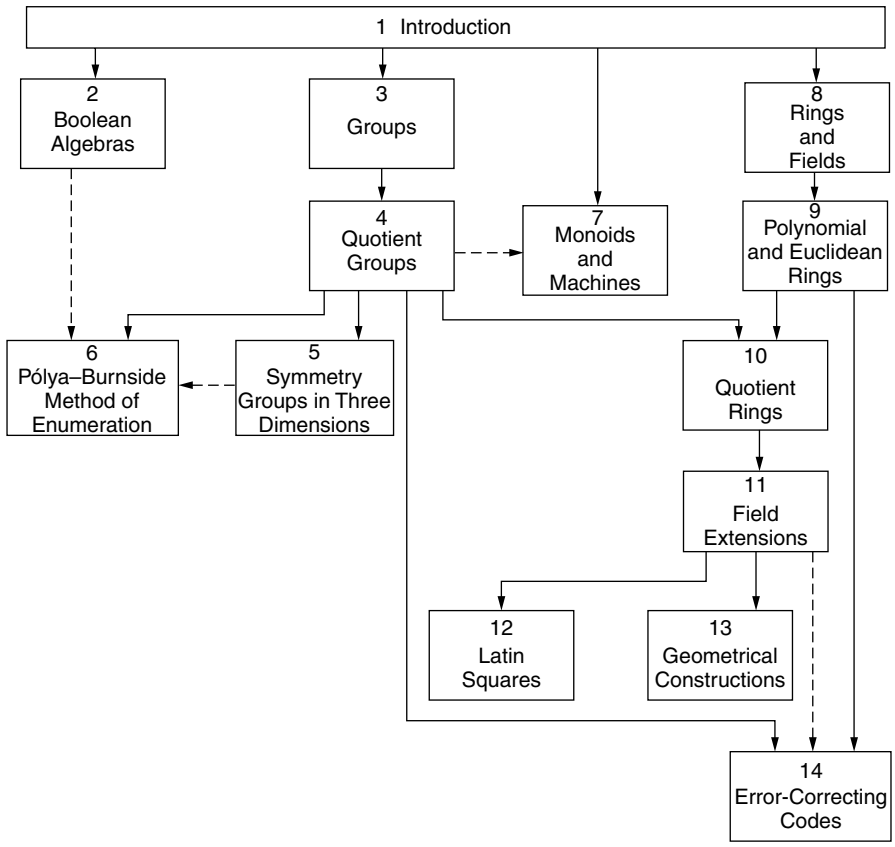


Figure P.1. Structure of the chapters.

The interested reader may delve further into any topic by consulting the books in the bibliography.

It is important to realize that the study of these applications is not the only reason for learning modern algebra. These examples illustrate the varied uses to which algebra has been put in the past, and it is extremely likely that many more different applications will be found in the future.

One cannot understand mathematics without doing numerous examples. There are a total of over 600 exercises of varying difficulty, at the ends of chapters. Answers to the odd-numbered exercises are given at the back of the book.

Figure P.1 illustrates the interdependence of the chapters. A solid line indicates a necessary prerequisite for the whole chapter, and a dashed line indicates a prerequisite for one section of the chapter. Since the book contains more than sufficient material for a two-term course, various sections or chapters may be omitted. The choice of topics will depend on the interests of the students and the instructor. However, to preserve the essence of the book, the instructor should be careful not to devote most of the course to the theory, but should leave sufficient time for the applications to be appreciated.

I would like to thank all my students and colleagues at the University of Waterloo, especially Harry Davis, D. Ž. Djoković, Denis Higgs, and Keith Rowe, who offered helpful suggestions during the various stages of the manuscript. I am very grateful to Michael Boyle, Ian McGee, Juris Stepŕans, and Jack Weiner for their help in preparing and proofreading the preliminary versions and the final draft. Finally, I would like to thank Sue Cooper, Annemarie DeBrusk, Lois Graham, and Denise Stack for their excellent typing of the different drafts, and Nadia Bahar for tracing all the figures.

Waterloo, Ontario, Canada WILLIAM J. GILBERT
April 1976

PREFACE TO THE SECOND EDITION

In addition to improvements in exposition, the second edition contains the following new items:

- New shorter proof of the parity theorem using the action of the symmetric group on the discriminant polynomial
- New proof that linear isometries are linear, and more detail about their relation to orthogonal matrices
- Appendix on methods of proof for beginning students, including the definition of an implication, proof by contradiction, converses, and logical equivalence
- Appendix on basic number theory covering induction, greatest common divisors, least common multiples, and the prime factorization theorem
- New material on the order of an element and cyclic groups
- More detail about the lattice of divisors of an integer
- New historical notes on Fermat's last theorem, the classification theorem for finite simple groups, finite affine planes, and more
- More detail on set theory and composition of functions
- 26 new exercises, 46 counting parts
- Updated symbols and notation
- Updated bibliography

February 2003

WILLIAM J. GILBERT
W. KEITH NICHOLSON

LIST OF SYMBOLS

$\mathbb{A}$	Algebraic numbers, 233
A_n	Alternating group on n elements, 70
$\mathbb{C}$	Complex numbers, 4
$\mathbb{C}^*$	Nonzero complex numbers, 48
C_n	Cyclic group of order n, 58
$C[0, \infty)$	Continuous real valued functions on $[0, \infty)$, 173
D_n	Dihedral group of order $2n$, 58
$\mathbb{D}_n$	Divisors of n, 22
$d(u, v)$	Hamming distance between u and v, 269
deg	Degree of a polynomial, 166
e	Identity element of a group or monoid, 48, 137
e_G	Identity element in the group G, 61
$\mathrm{E}(n)$	Euclidean group in n dimensions, 104
F	Field, 4, 160
$\mathcal{F}_n$	Switching functions of n variables, 28
$\mathrm{Fix}g$	Set of elements fixed under the action of g, 125
$\mathrm{FM}(A)$	Free monoid on A, 140
$\gcd(a, b)$	Greatest common divisor of a and b, 184, 299
$\mathrm{GF}(n)$	Galois field of order n, 227
$\mathrm{GL}(n, F)$	General linear group of dimension n over F, 107
$\mathbb{H}$	Quaternions, 177
I	Identity matrix, 4
I_k	$k \times k$ identity matrix, 277
$\mathrm{Im}\, f$	Image of f, 87
$\mathrm{Ker}\, f$	Kernel of f, 86
$\mathrm{lcm}(a, b)$	Least common multiple of a and b, 184, 303
$\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$	Linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^n$, 163
$M_n(R)$	$n \times n$ matrices with entries from R, 4, 166
$\mathbb{N}$	Nonnegative integers, 55
NAND	NOT-AND, 28, 36
NOR	NOT-OR, 28, 36
$\mathrm{O}(n)$	Orthogonal group of dimension n, 105
$\mathrm{Orb}\, x$	Orbit of x, 97

$\mathbb{P}$	Positive integers, 3
$\mathcal{P}(X)$	Power set of X, 8
$\mathbb{Q}$	Rational numbers, 6
$\mathbb{Q}^*$	Nonzero rational numbers, 48
Q	Quaternion group, 73
$\mathbb{R}$	Real numbers, 2
$\mathbb{R}^*$	Nonzero real numbers, 48
$\mathbb{R}^+$	Positive real numbers, 5
$S(X)$	Symmetric group of X, 50
S_n	Symmetric group on n elements, 63
$\mathrm{SO}(n)$	Special orthogonal group of dimension n, 108
Stab x	Stabilizer of x, 97
$\mathrm{SU}(n)$	Special unitary group of dimension n, 108
$\mathrm{T}(n)$	Translations in n dimensions, 104
$\mathrm{U}(n)$	Unitary group of dimension n, 108
$\mathbb{Z}$	Integers, 5
$\mathbb{Z}_n$	Integers modulo n, 5, 78
$\mathbb{Z}_n^*$	Integers modulo n coprime to n, 102
$\delta(x)$	Dirac delta function, or remainder in general division algorithm, 172, 181
Λ	Null sequence, 140
$\emptyset$	Empty set, 7
$\phi(n)$	Euler ϕ-function, 102
$\star$	General binary operation *or* concatenation, 2, 140
$*$	Convolution, 168, 173
$\circ$	Composition, 49
Δ	Symmetric difference, 9, 29
$-$	Difference, 9
$\wedge$	Meet, 14
$\vee$	Join, 14
$\subseteq$	Inclusion, 7
$\leqslant$	Less than or equal, 23
$\Rightarrow$	Implies, 17, 293
$\Leftrightarrow$	If and only if, 18, 295
$\cong$	Isomorphic, 60, 172
$\equiv \bmod n$	Congruent modulo n, 77
$\equiv \bmod H$	Congruent modulo H, 79
$\lvert X\rvert$	Number of elements in X, 12, 56
$\lvert G : H\rvert$	Index of H in G, 80
R^*	Invertible elements in the ring R, 188
a'	Complement of a in a boolean algebra, 14, 28
a^{-1}	Inverse of a, 3, 48
$\overline{A}$	Complement of the set A, 8
$\cap$	Intersection of sets, 8
$\cup$	Union of sets, 8

$\in$	Membership in a set, 7
$A-B$	Set difference, 9
$\|\mathbf{v}\|$	Length of $\mathbf{v}$ in $\mathbb{R}^n$, 105
$\mathbf{v} \cdot \mathbf{w}$	Inner product in $\mathbb{R}^n$, 105
V^T	Transpose of the matrix V, 104
$\square$	End of a proof or example, 9
(a)	Ideal generated by a, 204
$(a_1 a_2 \dots a_n)$	n-cycle, 64
$\begin{pmatrix} 1\ 2 \dots n \\ a_1 a_2 \dots a_n \end{pmatrix}$	Permutation, 63
$\binom{n}{r}$	Binomial coefficient $n!/r!(n-r)!$, 129
$F(a)$	Smallest field containing F and a, 220
$F(a_1, \dots, a_n)$	Smallest field containing F and $a_1, \dots, a_n$, 220
(n, k)-code	Code of length n with messages of length k, 266
$(X, \star)$	Group or monoid, 5, 48, 137
$(R, +, \cdot)$	Ring, 156
$(K, \wedge, \vee, ')$	Boolean algebra, 14
$[x]$	Equivalence class containing x, 77
$[x]_n$	Congruence class modulo n containing x, 100
$R[x]$	Polynomials in x with coefficients from R, 167
$R[[x]]$	Formal power series in x with coefficients from R, 169
$R[x_1, \dots, x_n]$	Polynomials in $x_1, \dots, x_n$ with coefficients from R, 168
$[K : F]$	Degree of K over F, 219
X^Y	Set of functions from Y to X, 138
$R^{\mathbb{N}}$	Sequences of elements from R, 168
$\langle a_i \rangle$	Sequence whose ith term is a_i, 168
$G \times H$	Direct product of G and H, 91
$S \times S$	Direct product of sets, 2
S/E	Quotient set, 77
G/H	Quotient group or set of right cosets, 83
R/I	Quotient ring, 206
$a\|b$	a divides b, 21, 184, 299
$l//m$	l is parallel to m, 242
Ha	Right coset of H containing a, 79
aH	Left coset of H containing a, 82
$I + r$	Coset of I containing r, 205

1

INTRODUCTION

Algebra can be defined as the manipulation of symbols. Its history falls into two distinct parts, with the dividing date being approximately 1800. The algebra done before the nineteenth century is called *classical algebra*, whereas most of that done later is called *modern algebra* or *abstract algebra*.

CLASSICAL ALGEBRA

The technique of introducing a symbol, such as x, to represent an unknown number in solving problems was known to the ancient Greeks. This symbol could be manipulated just like the arithmetic symbols until a solution was obtained. *Classical algebra* can be characterized by the fact that each symbol *always* stood for a number. This number could be integral, real, or complex. However, in the seventeenth and eighteenth centuries, mathematicians were not quite sure whether the square root of -1 was a number. It was not until the nineteenth century and the beginning of modern algebra that a satisfactory explanation of the complex numbers was given.

The main goal of classical algebra was to use algebraic manipulation to solve polynomial equations. Classical algebra succeeded in producing algorithms for solving all polynomial equations in one variable of degree at most four. However, it was shown by Niels Henrik Abel (1802–1829), by modern algebraic methods, that it was not always possible to solve a polynomial equation of degree five or higher in terms of nth roots. Classical algebra also developed methods for dealing with linear equations containing several variables, but little was known about the solution of nonlinear equations.

Classical algebra provided a powerful tool for tackling many scientific problems, and it is still extremely important today. Perhaps the most useful mathematical tool in science, engineering, and the social sciences is the method of solution of a system of linear equations together with all its allied linear algebra.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

MODERN ALGEBRA

In the nineteenth century it was gradually realized that mathematical symbols did not necessarily have to stand for numbers; in fact, it was not necessary that they stand for anything at all! From this realization emerged what is now known as *modern algebra* or *abstract algebra*.

For example, the symbols could be interpreted as symmetries of an object, as the position of a switch, as an instruction to a machine, or as a way to design a statistical experiment. The symbols could be manipulated using some of the usual rules for numbers. For example, the polynomial $3x^2 + 2x - 1$ could be added to and multiplied by other polynomials without ever having to interpret the symbol x as a number.

Modern algebra has two basic uses. The first is to describe patterns or symmetries that occur in nature and in mathematics. For example, it can describe the different crystal formations in which certain chemical substances are found and can be used to show the similarity between the logic of switching circuits and the algebra of subsets of a set. The second basic use of modern algebra is to extend the common number systems naturally to other useful systems.

BINARY OPERATIONS

The symbols that are to be manipulated are elements of some set, and the manipulation is done by performing certain operations on elements of that set. Examples of such operations are addition and multiplication on the set of real numbers.

As shown in Figure 1.1, we can visualize an operation as a "black box" with various inputs coming from a set S and one output, which combines the inputs in some specified way. If the black box has two inputs, the operation combines *two* elements of the set to form a third. Such an operation is called a *binary operation*. If there is only one input, the operation is called *unary*. An example of a unary operation is finding the reciprocal of a nonzero real number.

If S is a set, the **direct product** $S \times S$ consists of all ordered pairs (a, b) with $a, b \in S$. Here the term *ordered* means that $(a, b) = (a_1, b_1)$ if and only if $a = a_1$ and $b = b_1$. For example, if we denote the set of all real numbers by $\mathbb{R}$, then $\mathbb{R} \times \mathbb{R}$ is the euclidean plane.

Using this terminology, a **binary operation**, $\star$, on a set S is really just a particular function from $S \times S$ to S. We denote the image of the pair (a, b)

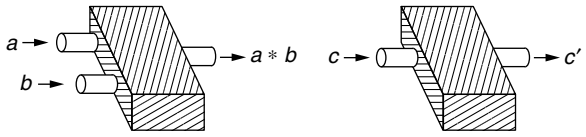


Figure 1.1

under this function by $a \star b$. In other words, the binary operation $\star$ assigns to any two elements a and b of S the element $a \star b$ of S. We often refer to an operation $\star$ as being **closed** to emphasize that each element $a \star b$ belongs to the set S and not to a possibly larger set. Many symbols are used for binary operations; the most common are $+$, $\cdot$, $-$, $\circ$, $\div$, $\cup$, $\cap$, $\wedge$, and $\vee$.

A **unary operation** on S is just a function from S to S. The image of c under a unary operation is usually denoted by a symbol such as c', $\overline{c}$, c^{-1}, or $(-c)$.

Let $\mathbb{P} = \{1, 2, 3, \ldots\}$ be the set of positive integers. Addition and multiplication are both binary operations on $\mathbb{P}$, because, if $x, y \in \mathbb{P}$, then $x + y$ and $x \cdot y \in \mathbb{P}$. However, subtraction is *not* a binary operation on $\mathbb{P}$ because, for instance, $1 - 2 \notin \mathbb{P}$. Other natural binary operations on $\mathbb{P}$ are exponentiation and the greatest common divisor, since for any two positive integers x and y, x^y and $\gcd(x, y)$ are well-defined elements of $\mathbb{P}$.

Addition, multiplication, and subtraction are all binary operations on $\mathbb{R}$ because $x + y$, $x \cdot y$, and $x - y$ are real numbers for every pair of real numbers x and y. The symbol $-$ stands for a binary operation when used in an expression such as $x - y$, but it stands for the unary operation of taking the negative when used in the expression $-x$. Division is not a binary operation on $\mathbb{R}$ because division by zero is undefined. However, division is a binary operation on $\mathbb{R} - \{0\}$, the set of nonzero real numbers.

A binary operation on a finite set can often be presented conveniently by means of a **table**. For example, consider the set $T = \{a, b, c\}$, containing three elements. A binary operation $\star$ on T is defined by Table 1.1. In this table, $x \star y$ is the element in row x and column y. For example, $b \star c = b$ and $c \star b = a$.

One important binary operation is the composition of symmetries of a given figure or object. Consider a square lying in a plane. The set S of symmetries of this square is the set of mappings of the square to itself that preserve distances. Figure 1.2 illustrates the composition of two such symmetries to form a third symmetry.

Most of the binary operations we use have one or more of the following special properties. Let $\star$ be a binary operation on a set S. This operation is called **associative** if $a \star (b \star c) = (a \star b) \star c$ for all $a, b, c \in S$. The operation $\star$ is called **commutative** if $a \star b = b \star a$ for all $a, b \in S$. The element $e \in S$ is said to be an **identity** for $\star$ if $a \star e = e \star a = a$ for all $a \in S$.

If $\star$ is a binary operation on S that has an identity e, then b is called the **inverse** of a with respect to $\star$ if $a \star b = b \star a = e$. We usually denote the

TABLE 1.1. Binary Operation on $\{a, b, c\}$

$\star$	a	b	c
a	b	a	a
b	c	a	b
c	c	a	b

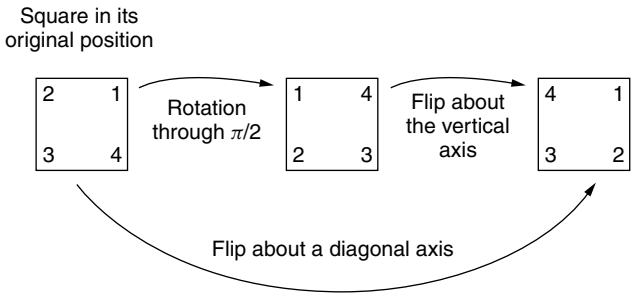


Figure 1.2. Composition of symmetries of a square.

inverse of a by a^{-1}; however, if the operation is addition, the inverse is denoted by $-a$.

If $\star$ and $\circ$ are two binary operations on S, then $\circ$ is said to be *distributive* over $\star$ if $a \circ (b \star c) = (a \circ b) \star (a \circ c)$ and $(b \star c) \circ a = (b \circ a) \star (c \circ a)$ for all $a, b, c \in S$.

Addition and multiplication are both associative and commutative operations on the set $\mathbb{R}$ of real numbers. The identity for addition is 0, whereas the multiplicative identity is 1. Every real number, a, has an inverse under addition, namely, its negative, $-a$. Every nonzero real number a has a multiplicative inverse, a^{-1}. Furthermore, multiplication is distributive over addition because $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ and $(b + c) \cdot a = (b \cdot a) + (c \cdot a)$; however, addition is not distributive over multiplication because $a + (b \cdot c) \neq (a + b) \cdot (a + c)$ in general.

Denote the set of $n \times n$ real matrices by $M_n(\mathbb{R})$. Matrix multiplication is an associative operation on $M_n(\mathbb{R})$, but it is not commutative (unless $n = 1$). The matrix I, whose (i, j)th entry is 1 if $i = j$ and 0 otherwise, is the multiplicative identity. Matrices with multiplicative inverses are called **nonsingular**.

ALGEBRAIC STRUCTURES

A set, together with one or more operations on the set, is called an **algebraic structure**. The set is called the **underlying set** of the structure. Modern algebra is the study of these structures; in later chapters, we examine various types of algebraic structures. For example, a field is an algebraic structure consisting of a set F together with two binary operations, usually denoted by $+$ and $\cdot$, that satisfy certain conditions. We denote such a structure by $(F, +, \cdot)$.

In order to understand a particular structure, we usually begin by examining its *substructures*. The underlying set of a substructure is a subset of the underlying set of the structure, and the operations in both structures are the same. For example, the set of complex numbers, $\mathbb{C}$, contains the set of real numbers, $\mathbb{R}$, as a subset. The operations of addition and multiplication on $\mathbb{C}$ restrict to the same operations on $\mathbb{R}$, and therefore $(\mathbb{R}, +, \cdot)$ is a substructure of $(\mathbb{C}, +, \cdot)$.

Two algebraic structures of a particular type may be compared by means of structure-preserving functions called *morphisms*. This concept of morphism is one of the fundamental notions of modern algebra. We encounter it among every algebraic structure we consider.

More precisely, let $(S, \star)$ and $(T, \circ)$ be two algebraic structures consisting of the sets S and T, together with the binary operations $\star$ on S and $\circ$ on T. Then a function $f: S \to T$ is said to be a **morphism** from $(S, \star)$ to $(T, \circ)$ if for every $x, y \in S$,

$$f(x \star y) = f(x) \circ f(y).$$

If the structures contain more than one operation, the morphism must preserve all these operations. Furthermore, if the structures have identities, these must be preserved, too.

As an example of a morphism, consider the set of all integers, $\mathbb{Z}$, under the operation of addition and the set of positive real numbers, $\mathbb{R}^+$, under multiplication. The function $f: \mathbb{Z} \to \mathbb{R}^+$ defined by $f(x) = e^x$ is a morphism from $(\mathbb{Z}, +)$ to $(\mathbb{R}^+, \cdot)$. Multiplication of the exponentials e^x and e^y corresponds to addition of their exponents x and y.

A *vector space* is an algebraic structure whose underlying set is a set of vectors. Its operations consist of the binary operation of addition and, for each scalar λ, a unary operation of multiplication by λ. A function $f: S \to T$, between vector spaces, is a morphism if $f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$ and $f(\lambda\mathbf{x}) = \lambda f(\mathbf{x})$ for all vectors $\mathbf{x}$ and $\mathbf{y}$ in the domain S and all scalars λ. Such a vector space morphism is usually called a **linear transformation**.

A morphism preserves some, but not necessarily all, of the properties of the domain structure. However, if a morphism between two structures is a bijective function (that is, one-to-one and onto), it is called an **isomorphism**, and the structures are called **isomorphic**. Isomorphic structures have identical properties, and they are indistinguishable from an algebraic point of view. For example, two vector spaces of the same finite dimension over a field F are isomorphic.

One important method of constructing new algebraic structures from old ones is by means of equivalence relations. If $(S, \star)$ is a structure consisting of the set S with the binary operation $\star$ on it, the equivalence relation $\sim$ on S is said to be *compatible* with $\star$ if, whenever $a \sim b$ and $c \sim d$, it follows that $a \star c \sim b \star d$. Such a compatible equivalence relation allows us to construct a new structure called the **quotient structure**, whose underlying set is the set of equivalence classes. For example, the quotient structure of the integers, $(\mathbb{Z}, +, \cdot)$, under the congruence relation modulo n, is the set of integers modulo n, $(\mathbb{Z}_n, +, \cdot)$ (see Appendix 2).

EXTENDING NUMBER SYSTEMS

In the words of Leopold Kronecker (1823–1891), "God created the natural numbers; everything else was man's handiwork." Starting with the set of natural

numbers under addition and multiplication, we show how this can be extended to other algebraic systems that satisfy properties not held by the natural numbers. The integers $(\mathbb{Z}, +, \cdot)$ is the smallest system containing the natural numbers, in which addition has an identity (the zero) and every element has an inverse under addition (its negative). The integers have an identity under multiplication (the element 1), but 1 and -1 are the only elements with multiplicative inverses. A standard construction will produce the field of fractions of the integers, which is the rational number system $(\mathbb{Q}, +, \cdot)$, and we show that this is the smallest field containing $(\mathbb{Z}, +, \cdot)$. We can now divide by nonzero elements in $\mathbb{Q}$ and solve every linear equation of the form $ax = b$ $(a \neq 0)$. However, not all quadratic equations have solutions in $\mathbb{Q}$; for example, $x^2 - 2 = 0$ has no rational solution.

The next step is to extend the rationals to the real number system $(\mathbb{R}, +, \cdot)$. The construction of the real numbers requires the use of nonalgebraic concepts such as Dedekind cuts or Cauchy sequences, and we will not pursue this, being content to assume that they have been constructed. Even though many polynomial equations have real solutions, there are some, such as $x^2 + 1 = 0$, that do not. We show how to extend the real number system by adjoining a root of $x^2 + 1$ to obtain the complex number system $(\mathbb{C}, +, \cdot)$. The complex number system is really the end of the line, because Carl Friedrich Gauss (1777–1855), in his doctoral thesis, proved that any nonconstant polynomial with real or complex coefficients has a root in the complex numbers. This result is now known as the *fundamental theorem of algebra*.

However, the classical number system can be generalized in a different way. We can look for fields that are not subfields of $(\mathbb{C}, +, \cdot)$. An example of such a field is the system of integers modulo a prime p, $(\mathbb{Z}_p, +, \cdot)$. All the usual operations of addition, subtraction, multiplication, and division by nonzero elements can be performed in $\mathbb{Z}_p$. We show that these fields can be extended and that for each prime p and positive integer n, there is a field $(GF(p^n), +, \cdot)$ with p^n elements. These finite fields are called **Galois fields** after the French mathematician Évariste Galois. We use Galois fields in the construction of orthogonal latin squares and in coding theory.

2

BOOLEAN ALGEBRAS

A boolean algebra is a good example of a type of algebraic structure in which the symbols usually represent nonnumerical objects. This algebra is modeled after the algebra of subsets of a set under the binary operations of union and intersection and the unary operation of complementation. However, boolean algebra has important applications to switching circuits, where each symbol represents a particular electrical circuit or switch. The origin of boolean algebra dates back to 1847, when the English mathematician George Boole (1815–1864) published a slim volume entitled *The Mathematical Analysis of Logic*, which showed how algebraic symbols could be applied to logic. The manipulation of logical propositions by means of boolean algebra is now called the *propositional calculus*.

At the end of this chapter, we show that any finite boolean algebra is equivalent to the algebra of subsets of a set; in other words, there is a boolean algebra isomorphism between the two algebras.

ALGEBRA OF SETS

In this section, we develop some properties of the basic operations on sets. A set is often referred to informally as a collection of objects called the elements of the set. This is not a proper definition—*collection* is just another word for *set*. What is clear is that there are **sets**, and there is a notion of being an **element** (or member) of a set. These fundamental ideas are the primitive concepts of set theory and are left undefined.* The fact that a is an element of a set X is denoted $a \in X$. If every element of X is also an element of Y, we write $X \subseteq Y$ (equivalently, $Y \supseteq X$) and say that X is *contained* in Y, or that X is a **subset** of Y. If X and Y have the same elements, we say that X and Y are **equal sets** and write $X = Y$. Hence $X = Y$ if and only if both $X \subseteq Y$ and $Y \subseteq X$. The set with no elements is called the **empty set** and is denoted as $\emptyset$.

* Certain basic properties of sets must also be assumed (called the *axioms* of the theory), but it is not our intention to go into this here.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

Let X be any set. The set of all subsets of X is called the **power set** of X and is denoted by $\mathcal{P}(X)$. Hence $\mathcal{P}(X) = \{A | A \subseteq X\}$. Thus if $X = \{a, b\}$, then $\mathcal{P}(X) = \{\emptyset, \{a\}, \{b\}, X\}$. If $X = \{1, 2, 3\}$, then $\mathcal{P}(X) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, X\}$.

If A and B are subsets of a set X, their **intersection** $A \cap B$ is defined to be the set of elements common to A and B, and their **union** $A \cup B$ is the set of elements in A or B (or both). More formally,

$$A \cap B = \{x | x \in A \text{ and } x \in B\} \quad \text{and} \quad A \cup B = \{x | x \in A \text{ or } x \in B\}.$$

The **complement** of A in X is $\overline{A} = \{x | x \in X \text{ and } x \notin A\}$ and is the set of elements in X that are not in A. The shaded areas of the **Venn diagrams** in Figure 2.1 illustrate these operations.

Union and intersection are both **binary operations** on the power set $\mathcal{P}(X)$, whereas complementation is a **unary operation** on $\mathcal{P}(X)$. For example, with $X = \{a, b\}$, the tables for the structures $(\mathcal{P}(X), \cap)$, $(\mathcal{P}(X), \cup)$ and $(\mathcal{P}(X), {}^-)$ are given in Table 2.1, where we write A for $\{a\}$ and B for $\{b\}$.

Proposition 2.1. The following are some of the more important relations involving the operations $\cap$, $\cup$, and $^-$, holding for all $A, B, C \in \mathcal{P}(X)$.

(i) $A \cap (B \cap C) = (A \cap B) \cap C$.
(ii) $A \cup (B \cup C) = (A \cup B) \cup C$.
(iii) $A \cap B = B \cap A$.
(iv) $A \cup B = B \cup A$.
(v) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
(vi) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
(vii) $A \cap X = A$.
(viii) $A \cup \emptyset = A$.
(ix) $A \cap \overline{A} = \emptyset$.
(x) $A \cup \overline{A} = X$.
(xi) $A \cap \emptyset = \emptyset$.
(xii) $A \cup X = X$.
(xiii) $A \cap (A \cup B) = A$.
(xiv) $A \cup (A \cap B) = A$.

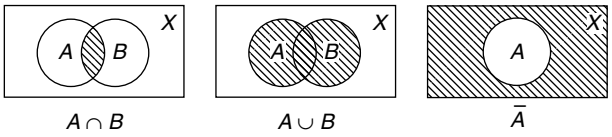


Figure 2.1. Venn diagrams.

TABLE 2.1. Intersection, Union, and Complements in $\mathcal{P}(\{a, b\})$

$\cap$	$\emptyset$	A	B	X
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
A	$\emptyset$	A	$\emptyset$	A
B	$\emptyset$	$\emptyset$	B	B
X	$\emptyset$	A	B	X

$\cup$	$\emptyset$	A	B	X
$\emptyset$	$\emptyset$	A	B	X
A	A	A	X	X
B	B	X	B	X
X	X	X	X	X

Subset	Complement
$\emptyset$	X
A	B
B	A
X	$\emptyset$

(xv) $A \cap A = A$.
(xvi) $A \cup A = A$.
(xvii) $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$.
(xviii) $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$.
(xix) $\overline{X} = \emptyset$.
(xx) $\overline{\emptyset} = X$.
(xxi) $\overline{\overline{A}} = A$.

Proof. We shall prove relations (v) and (x) and leave the proofs of the others to the reader.

$$\begin{aligned}
\text{(v)}\quad A \cap (B \cup C) &= \{x | x \in A \text{ and } x \in B \cup C\} \\
&= \{x | x \in A \text{ and } (x \in B \text{ or } x \in C)\} \\
&= \{x | (x \in A \text{ and } x \in B) \text{ or } (x \in A \text{ and } x \in C)\} \\
&= \{x | x \in A \cap B \text{ or } x \in A \cap C\} \\
&= (A \cap B) \cup (A \cap C).
\end{aligned}$$

The Venn diagrams in Figure 2.2 illustrate this result.

$$\begin{aligned}
\text{(x)}\quad A \cup \overline{A} &= \{x | x \in A \text{ or } x \in \overline{A}\} \\
&= \{x | x \in A \text{ or } (x \in X \text{ and } x \notin A)\} \\
&= \{x | (x \in X \text{ and } x \in A) \text{ or } (x \in X \text{ and } x \notin A)\}, \text{ since } A \subseteq X \\
&= \{x | x \in X \text{ and } (x \in A \text{ or } x \notin A)\} \\
&= \{x | x \in X\}, \text{ since it is always true that } x \in A \text{ or } x \notin A \\
&= X. \qquad \square
\end{aligned}$$

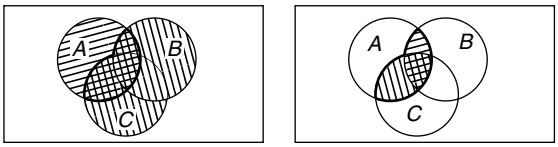
Relations (i)–(iv), (vii), and (viii) show that $\cap$ and $\cup$ are associative and commutative operations on $\mathcal{P}(X)$ with identities X and $\emptyset$, respectively. The only element with an inverse under $\cap$ is its identity X, and the only element with an inverse under $\cup$ is its identity $\emptyset$.

Note the duality between $\cap$ and $\cup$. If these operations are interchanged in any relation, the resulting relation is also true.

Another operation on $\mathcal{P}(X)$ is the **difference** of two subsets. It is defined by

$$A - B = \{x | x \in A \text{ and } x \notin B\} = A \cap \overline{B}.$$

Since this operation is neither associative nor commutative, we introduce another operation $A \Delta B$, called the **symmetric difference**, illustrated in Figure 2.3,



$A \cap (B \cup C)$ $(A \cap B) \cup (A \cap C)$

Figure 2.2. Venn diagrams illustrating a distributive law.

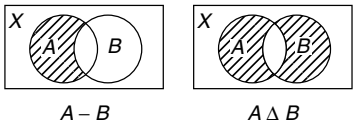


Figure 2.3. Difference and symmetric difference of sets.

defined by

$$A\Delta B = (A \cap \overline{B}) \cup (\overline{A} \cap B) = (A \cup B) - (A \cap B) = (A - B) \cup (B - A).$$

The symmetric difference of A and B is the set of elements in A or B, but not in both. This is often referred to as the **exclusive OR function** of A and B.

Example 2.2. Write down the table for the structure $(\mathcal{P}(X), \Delta)$ when $X = \{a, b\}$.

Solution. The table is given in Table 2.2, where we write A for $\{a\}$ and B for $\{b\}$. □

Proposition 2.3. The operation Δ is associative and commutative on $\mathcal{P}(X)$; it has an identity $\emptyset$, and each element is its own inverse. That is, the following relations hold for all $A, B, C \in \mathcal{P}(X)$:

(i) $A\Delta(B\Delta C) = (A\Delta B)\Delta C$. (ii) $A\Delta B = B\Delta A$.
(iii) $A\Delta\emptyset = A$. (iv) $A\Delta A = \emptyset$.

Three further properties of the symmetric difference are:

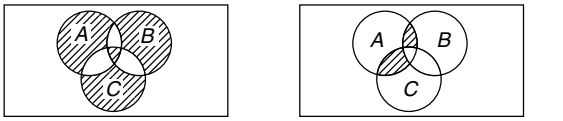
(v) $A\Delta X = \overline{A}$. (vi) $A\Delta\overline{A} = X$.
(vii) $A \cap (B\Delta C) = (A \cap B)\Delta(A \cap C)$.

Proof. (ii) follows because the definition of $A\Delta B$ is symmetric in A and B. To prove (i) observe first that Proposition 2.1 gives

$$\begin{aligned}\overline{B\Delta C} &= \overline{(B \cap \overline{C}) \cup (\overline{B} \cap C)} = (\overline{B} \cup C) \cap (B \cup \overline{C}) \\ &= (\overline{B} \cap B) \cup (\overline{B} \cap \overline{C}) \cup (C \cap B) \cup (C \cap \overline{C}) \\ &= (B \cap C) \cup (\overline{B} \cap \overline{C}).\end{aligned}$$

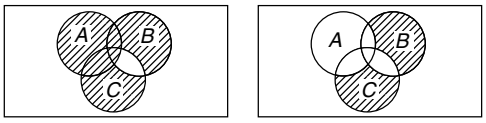
TABLE 2.2. Symmetric Difference in $\mathcal{P}(\{a, b\})$

Δ	$\emptyset$	A	B	X
$\emptyset$	$\emptyset$	A	B	X
A	A	$\emptyset$	X	B
B	B	X	$\emptyset$	A
X	X	B	A	$\emptyset$



$A \Delta (B \Delta C) = (A \Delta B) \Delta C$ $A \cap (B \Delta C) = (A \cap B) \Delta (A \cap C)$

Figure 2.4. Venn diagrams.



$A \cup (B \Delta C)$ $(A \cup B) \Delta (A \cup C)$

Figure 2.5. Venn diagrams of unequal expressions.

Hence

$$\begin{aligned} A\Delta(B\Delta C) &= \{A \cap (\overline{B\Delta C})\} \cup \{\overline{A} \cap (B\Delta C)\} \\ &= \{A \cap [(B \cap C) \cup (\overline{B} \cap \overline{C})]\} \cup \{\overline{A} \cap [(B \cap \overline{C}) \cup (\overline{B} \cap C)]\} \\ &= (A \cap B \cap C) \cup (A \cap \overline{B} \cap \overline{C}) \cup (\overline{A} \cap B \cap \overline{C}) \cup (\overline{A} \cap \overline{B} \cap C)). \end{aligned}$$

This expression is symmetric in A, B, and C, so (ii) gives

$$A\Delta(B\Delta C) = C\Delta(A\Delta B) = (A\Delta B)\Delta C.$$

We leave the proof of the other parts to the reader. Parts (i) and (vii) are illustrated in Figure 2.4. □

Relation (vii) of Proposition 2.3 is a distributive law and states that $\cap$ is distributive over Δ. It is natural to ask whether $\cup$ is distributive over Δ.

Example 2.4. Is it true that $A \cup (B\Delta C) = (A \cup B)\Delta(A \cup C)$ for all $A, B, C \in \mathcal{P}(X)$?

Solution. The Venn diagrams for each side of the equation are given in Figure 2.5. If the shaded areas are not the same, we will be able to find a counter example. We see from the diagrams that the result will be false if A is nonempty. If $A = X$ and $B = C = \emptyset$, then $A \cup (B\Delta C) = A$, whereas $(A \cup B)\Delta(A \cup C) = \emptyset$; thus union is not distributive over symmetric difference. □

NUMBER OF ELEMENTS IN A SET

If a set X contains two or three elements, we have seen that $\mathcal{P}(X)$ contains 2^2 or 2^3 elements, respectively. This suggests the following general result on the number of subsets of a finite set.

Theorem 2.5. If X is a finite set with n elements, then $\mathcal{P}(X)$ contains 2^n elements.

Proof. Each of the n elements of X is either in a given subset A or not in A. Hence, in choosing a subset of X, we have two choices for each element, and these choices are independent. Therefore, the number of choices is 2^n, and this is the number of subsets of X.

If $n = 0$, then $X = \emptyset$ and $\mathcal{P}(X) = \{\emptyset\}$, which contains one element. □

Denote the number of elements of a set X by $|X|$. If A and B are finite **disjoint sets** (that is, $A \cap B = \emptyset$), then

$$|A \cup B| = |A| + |B|.$$

Proposition 2.6. For any two finite sets A and B,

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Proof. We can express $A \cup B$ as the disjoint union of A and $B - A$; also, B can be expressed as the disjoint union of $B - A$ and $A \cap B$ as shown in Figure 2.6. Hence $|A \cup B| = |A| + |B - A|$ and $|B| = |B - A| + |A \cap B|$. It follows that $|A \cup B| = |A| + |B| - |A \cap B|$. □

Proposition 2.7. For any three finite sets A, B, and C,

$$\begin{aligned} |A \cup B \cup C| = {} & |A| + |B| + |C| - |A \cap B| - |A \cap C| \\ & - |B \cap C| + |A \cap B \cap C|. \end{aligned}$$

Proof. Write $A \cup B \cup C$ as $(A \cup B) \cup C$. Then, by Proposition 2.6,

$$\begin{aligned} |A \cup B \cup C| &= |A \cup B| + |C| - |(A \cup B) \cap C| \\ &= |A| + |B| - |A \cap B| + |C| - |(A \cap C) \cup (B \cap C)| \\ &= |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| \\ &\quad + |(A \cap C) \cap (B \cap C)|. \end{aligned}$$

The result follows because $(A \cap C) \cap (B \cap C) = A \cap B \cap C$. □

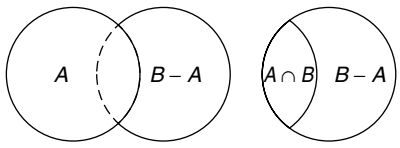


Figure 2.6

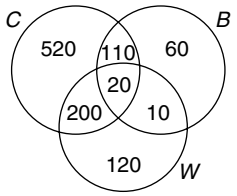


Figure 2.7. Different classes of commuters.

Example 2.8. A survey of 1000 commuters reported that 850 sometimes used a car, 200 a bicycle, and 350 walked, whereas 130 used a car and a bicycle, 220 used a car and walked, 30 used a bicycle and walked, and 20 used all three. Are these figures consistent?

Solution. Let C, B, and W be the sets of commuters who sometimes used a car, a bicycle, and walked, respectively. Then

$$\begin{aligned} |C \cup B \cup W| &= |C| + |B| + |W| - |C \cap B| - |C \cap W| \\ &\quad - |B \cap W| + |C \cap B \cap W| \\ &= 850 + 200 + 350 - 130 - 220 - 30 + 20 \\ &= 1040. \end{aligned}$$

Since this number is greater than 1000, the figures must be inconsistent. The breakdown of the reported figures into their various classes is illustrated in Figure 2.7. The sum of all these numbers is 1040. □

Example 2.9. If 47% of the people in a community voted in a local election and 75% voted in a federal election, what is the least percentage that voted in both?

Solution. Let L and F be the sets of people who voted in the local and federal elections, respectively. If n is the total number of voters in the community, then $|L| + |F| - |L \cap F| = |L \cup F| \leqslant n$. It follows that

$$|L \cap F| \geqslant |L| + |F| - n = \left(\frac{47}{100} + \frac{75}{100} - 1\right) n = \frac{22}{100} n.$$

Hence at least 22% voted in both elections. □

BOOLEAN ALGEBRAS

We now give the definition of an abstract boolean algebra in terms of a set with two binary operations and one unary operation on it. We show that various algebraic structures, such as the algebra of sets, the logic of propositions, and

the algebra of switching circuits are all boolean algebras. It then follows that any general result derived from the axioms will hold in all our examples of boolean algebras.

It should be noted that this axiom system is only one of many equivalent ways of defining a boolean algebra. Another common way is to define a boolean algebra as a lattice satisfying certain properties (see the section "Posets and Lattices").

A **boolean algebra** $(K, \wedge, \vee, ')$ is a set K together with two binary operations $\wedge$ and $\vee$, and a unary operation $'$ on K satisfying the following axioms for all $A, B, C \in K$:

(i) $A \wedge (B \wedge C) = (A \wedge B) \wedge C$.
(ii) $A \vee (B \vee C) = (A \vee B) \vee C$.
(associative laws)
(iii) $A \wedge B = B \wedge A$.
(iv) $A \vee B = B \vee A$.
(commutative laws)
(v) $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$.
(vi) $A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C)$.
(distributive laws)
(vii) There is a **zero element** 0 in K such that $A \vee 0 = A$.
(viii) There is a **unit element** 1 in K such that $A \wedge 1 = A$.
(ix) $A \wedge A' = 0$.
(x) $A \vee A' = 1$.

We call the operations $\wedge$ and $\vee$, **meet** and **join**, respectively. The element A' is called the **complement** of A.

The associative axioms (i) and (ii) are redundant in the system above because with a little effort they can be deduced from the other axioms. However, since associativity is such an important property, we keep these properties as axioms.

It follows from Proposition 2.1 that $(\mathcal{P}(X), \cap, \cup, {}^-)$ is a boolean algebra with $\emptyset$ as zero and X as unit. When $X = \emptyset$, this boolean algebra of subsets contains one element, and this is both the zero and unit. It can be proved (see Exercise 2.17) that if the zero and unit elements are the same, the boolean algebra must have only one element.

We can define a two-element boolean algebra $(\{0, 1\}, \wedge, \vee, ')$ by means of Table 2.3.

Proposition 2.10. If the binary operation $\star$ on the set K has an identity e such that $a \star e = e \star a = a$ for all $a \in K$, then this identity is unique.

TABLE 2.3. Two-Element Boolean Algebra

A	B	$A \wedge B$	$A \vee B$
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	1

A	A'
0	1
1	0

Proof. Suppose that e and e' are both identities. Then $e = e \star e'$, since e' is an identity, and $e \star e' = e'$ since e is an identity. Hence $e = e'$, so the identity must be unique. □

Corollary 2.11. The zero and unit elements in a boolean algebra are unique.

Proof. This follows directly from the proposition above, because the zero and unit elements are the identities for the join and meet operations, respectively. □

Proposition 2.12. The complement of an element in a boolean algebra is unique; that is, for each $A \in K$ there is only one element $A' \in K$ satisfying axioms (ix) and (x): $A \wedge A' = 0$ and $A \vee A' = 1$.

Proof. Suppose that B and C are both complements of A, so that $A \wedge B = 0$, $A \vee B = 1$, $A \wedge C = 0$, and $A \vee C = 1$. Then

$$\begin{aligned} B &= B \vee 0 = B \vee (A \wedge C) = (B \vee A) \wedge (B \vee C) \\ &= (A \vee B) \wedge (B \vee C) = 1 \wedge (B \vee C) = B \vee C. \end{aligned}$$

Similarly, $C = C \vee B$ and so $B = B \vee C = C \vee B = C$. □

If we interchange $\wedge$ and $\vee$ and interchange 0 and 1 in the system of axioms for a boolean algebra, we obtain the same system. Therefore, if any proposition is derivable from the axioms, so is the proposition obtained by interchanging $\wedge$ and $\vee$ and interchanging 0 and 1. This is called the **duality principle**. For example, in the following proposition, there are four pairs of dual statements. If one member of each pair can be proved, the other will follow directly from the duality principle.

If $(K, \wedge, \vee, ')$ is a boolean algebra with 0 as zero and 1 as unit, then $(K, \vee, \wedge, ')$ is also a boolean algebra with 1 as zero and 0 as unit.

Proposition 2.13. If A, B, and C are elements of a boolean algebra $(K, \wedge, \vee, ')$, the following relations hold:

(i) $A \wedge 0 = 0$.
(ii) $A \vee 1 = 1$.
(iii) $A \wedge (A \vee B) = A$.
(iv) $A \vee (A \wedge B) = A$. (absorption laws)
(v) $A \wedge A = A$.
(vi) $A \vee A = A$. (idempotent laws)
(vii) $(A \wedge B)' = A' \vee B'$.
(viii) $(A \vee B)' = A' \wedge B'$. (De Morgan's laws)
(ix) $(A')' = A$.

Proof. Note first that relations (ii), (iv), (vi), and (viii) are the duals of relations (i), (iii), (v), and (vii), so we prove the last four, and relation (ix). We use the axioms for a boolean algebra several times.

(i) $A \wedge 0 = A \wedge (A \wedge A') = (A \wedge A) \wedge A' = A \wedge A' = 0.$

(iii) $A \wedge (A \vee B) = (A \vee 0) \wedge (A \vee B) = A \vee (0 \wedge B) = A \vee 0 = A.$

(v) $A = A \wedge 1 = A \wedge (A \vee A') = (A \wedge A) \vee (A \wedge A')$
$= (A \wedge A) \vee 0 = A \wedge A.$

Relations (vii) follows from Proposition 2.12 if we can show that $A' \vee B'$ is a complement of $A \wedge B$ [then it is *the* complement $(A \wedge B)'$]. Now using part (i) of this proposition,

$$\begin{aligned}(A \wedge B) \wedge (A' \vee B') &= [(A \wedge B) \wedge A'] \vee [(A \wedge B) \wedge B'] \\ &= [(A \wedge A') \wedge B] \vee [A \wedge (B \wedge B')] \\ &= [0 \wedge B] \vee [A \wedge 0] \\ &= 0 \vee 0 \\ &= 0.\end{aligned}$$

Similarly, part (ii) gives

$$\begin{aligned}(A \wedge B) \vee (A' \vee B') &= [A \vee (A' \vee B')] \wedge [B \vee (A' \vee B')] \\ &= [(A \vee A') \vee B'] \wedge [(B \vee B') \vee A'] \\ &= [1 \vee B'] \wedge [1 \vee A'] \\ &= 1 \wedge 1 \\ &= 1.\end{aligned}$$

To prove relation (ix), by definition we have $A' \wedge A = 0$ and $A' \vee A = 1$. Therefore, A is a complement of A', and since the complement is unique, $A = (A')'$. □

PROPOSITIONAL LOGIC

We now show briefly how boolean algebra can be applied to the logic of propositions. Consider two sentences "A" and "B", which may either be true or false. For example, "A" could be "This apple is red," and "B" could be "This pear is green." We can combine these to form other sentences, such as "A and B," which would be "This apple is red, and this pear is green." We could also form the sentence "not A," which would be "This apple is not red." Let us now compare the truth or falsity of the derived sentences with the truth or falsity of the original ones. We illustrate the relationship by means of a diagram called a **truth table**. Table 2.4 shows the truth tables for the expressions "A and B," "A or B," and "not A." In these tables, T stands for "true" and F stands for "false." For

TABLE 2.4. Truth Tables

A	B	A and B	A or B
F	F	F	F
F	T	F	T
T	F	F	T
T	T	T	T

A	Not A
F	T
T	F

example, if the statement "A" is true while "B" is false, the statement "A and B" will be false, and the statement "A or B" will be true.

We can have two seemingly different sentences with the same meaning; for example, "This apple is not red or this pear is not green" has the same meaning as "It is not true that this apple is red and that this pear is green." If two sentences, P and Q, have the same meaning, we say that P and Q are **logically equivalent**, and we write $P = Q$. The example above concerning apples and pears implies that

$$(\text{not } A) \text{ or } (\text{not } B) = \text{not } (A \text{ and } B).$$

This equation corresponds to De Morgan's law in a boolean algebra.

It appears that a set of sentences behaves like a boolean algebra. To be more precise, let us consider a set of sentences that are closed under the operations of "and," "or," and "not." Let K be the set, each element of which consists of all the sentences that are logically equivalent to a particular sentence. Then it can be verified that (K, and, or, not) is indeed a boolean algebra. The zero element is called a **contradiction**, that is, a statement that is always false, such as "This apple is red and this apple is not red." The unit element is called a **tautology**, that is, a statement that is always true, such as "This apple is red or this apple is not red." This allows us to manipulate logical propositions using formulas derived from the axioms of a boolean algebra.

An important method of combining two statements, A and B, in a sentence is by a **conditional**, such as "If A, then B," or equivalently, "A implies B," which we shall write as "$A \Rightarrow B$." How does the truth or falsity of such a conditional depend on that of A and B? Consider the following sentences:

1. If $x > 4$, then $x^2 > 16$.
2. If $x > 4$, then $x^2 = 2$.
3. If $2 = 3$, then $0.2 = 0.3$.
4. If $2 = 3$, then the moon is made of green cheese.

Clearly, if A is true, then B must also be true for the sentence "$A \Rightarrow B$" to be true. However, if A is not true, then the sentence "If A, then B" has no standard meaning in everyday language. Let us take "$A \Rightarrow B$" to mean that we cannot have A true and B not true. This implies that the truth value of the

TABLE 2.5. Truth tables for Conditional and Biconditional Statements

A	B	$A \Rightarrow B$	$A \Leftarrow B$	$A \Leftrightarrow B$
F	F	T	T	T
F	T	T	F	F
T	F	F	T	F
T	T	T	T	T

statement "$A \Rightarrow B$" is the same as that of "not (A and not B)." Let us write $\wedge$, $\vee$, and $'$ for "and," "or," and "not," respectively. Then "$A \Rightarrow B$" is equivalent to $(A \wedge B')' = A' \vee B$. Thus "$A \Rightarrow B$" is true if A is false or if B is true. Using this definition, statements 1, 3, and 4 are all true, whereas statement 2 is false.

We can combine two conditional statements to form a **biconditional statement** of the form "A if and only if B" or "$A \Leftrightarrow B$." This has the same truth value as "$(A \Rightarrow B)$ and $(B \Rightarrow A)$" or, equivalently, $(A \wedge B) \vee (A' \wedge B')$. Another way of expressing this biconditional is to say that "A is a necessary and sufficient condition for B." It is seen from Table 2.5 that the statement "$A \Leftrightarrow B$" is true if either A and B are both true or A and B are both false.

Example 2.14. Apply this propositional calculus to determine whether a certain politician's arguments are consistent. In one speech he states that if taxes are raised, the rate of inflation will drop if and only if the value of the dollar does not fall. On television, he says that if the rate of inflation decreases or the value of the dollar does not fall, taxes will not be raised. In a speech abroad, he states that either taxes must be raised or the value of the dollar will fall and the rate of inflation will decrease. His conclusion is that taxes will be raised, but the rate of inflation will decrease, and the value of the dollar will not fall.

Solution. We write

A to mean "Taxes will be raised,"
B to mean "The rate of inflation will decrease,"
C to mean "The value of the dollar will not fall."

The politician's three statements can be written symbolically as

(i) $A \Rightarrow (B \Leftrightarrow C)$.
(ii) $(B \vee C) \Rightarrow A'$.
(iii) $A \vee (C' \wedge B)$.

His conclusion is (iv) $A \wedge B \wedge C$.

The truth values of the first two statements are equivalent to those of the following:

TABLE 2.6. Truth Tables for the Politician's Arguments

A	B	C	(i)	(ii)	(iii)	(i) $\wedge$ (ii) $\wedge$ (iii)	(iv)	(i) $\wedge$ (ii) $\wedge$ (iii) $\Rightarrow$ (iv)
F	F	F	T	T	F	F	F	T
F	F	T	T	T	F	F	F	T
F	T	F	T	T	T	T	F	F
F	T	T	T	T	F	F	F	T
T	F	F	T	T	T	T	F	F
T	F	T	F	F	T	F	F	T
T	T	F	F	F	T	F	F	T
T	T	T	T	F	T	F	T	T

(i) $A' \vee ((B \wedge C) \vee (B' \wedge C'))$.

(ii) $(B \vee C)' \vee A'$.

It follows from Table 2.6 that (i) $\wedge$ (ii) $\wedge$ (iii) $\Rightarrow$ (iv) is not a tautology; that is, it is not always true. Therefore, the politician's arguments are incorrect. They break down when A and C are false and B is true, and when B and C are false and A is true. □

SWITCHING CIRCUITS

In this section we use boolean algebra to analyze some simple switching circuits. A **switch** is a device with two states; state 1 is the "on" state, and state 0 the "off" state. An ordinary household light switch is such a device, but the theory holds equally well for more sophisticated electronic or magnetic two-state devices. We analyze circuits with two terminals: The circuit is said to be **closed** if current can pass between the terminals, and **open** if current cannot pass.

We denote a switch A by the symbol in Figure 2.8. We assign the value 1 to A if the switch A is closed and the value 0 if it is open. We denote two switches by the same letter if they open and close simultaneously. If B is a switch that is always in the opposite position to A (that is, if B is open when A is closed, and B is closed when A is open), denote switch B by A'.

The two switches A and B in Figure 2.9 are said to be connected in **series**. If we connect this circuit to a power source and a light as in Figure 2.10, we see that the light will be on if and only if A *and* B are both switched on; we denote this series circuit by $A \wedge B$. Its effect is shown in Table 2.7.

The switches A and B in Figure 2.11 are said to be in **parallel**, and this circuit is denoted by $A \vee B$ because the circuit is closed if either A *or* B is switched on.

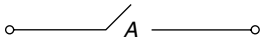


Figure 2.8. Switch A.



Figure 2.9. Switches in series.

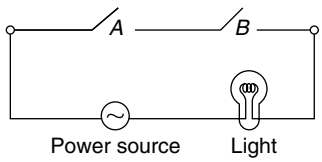


Figure 2.10. Series circuit.

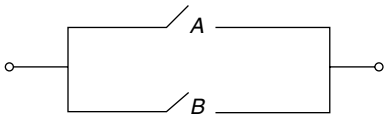


Figure 2.11. Switches in parallel.

TABLE 2.7. Effect of the Series Circuit

Switch A	Switch B	Circuit $A \wedge B$	Light
0 (off)	0 (off)	0 (open)	off
0 (off)	1 (on)	0 (open)	off
1 (on)	0 (off)	0 (open)	off
1 (on)	1 (on)	1 (closed)	on

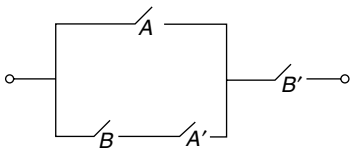


Figure 2.12. Series-parallel circuit.

The reader should be aware that many books on switching theory use the notation $+$ and $\cdot$ instead of $\vee$ and $\wedge$, respectively.

Series and parallel circuits can be combined to form circuits like the one in Figure 2.12. This circuit would be denoted by $(A \vee (B \wedge A')) \wedge B'$. Such circuits are called **series-parallel switching circuits**.

In actual practice, the wiring diagram may not look at all like Figure 2.12, because we would want switches A and A' together and B and B' together. Figure 2.13 illustrates one particular form that the wiring diagram could take.

Two circuits $\mathcal{C}_1$ and $\mathcal{C}_2$ involving the switches $A, B, \ldots$ are said to be **equivalent** if the positions of the switches $A, B, \ldots$, which allow current to pass,

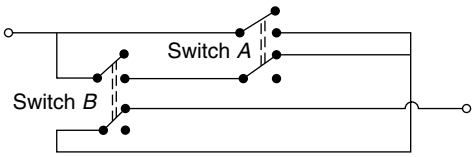


Figure 2.13. Wiring diagram of the circuit.

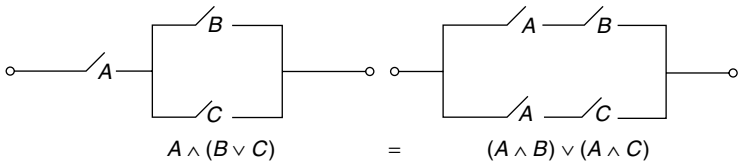


Figure 2.14. Distributive law.

are the same for both circuits. We write $\mathcal{C}_1 = \mathcal{C}_2$ to mean that the circuits are equivalent. It can be verified that all the axioms for a boolean algebra are valid when interpreted as series-parallel switching circuits. For example, Figure 2.14 illustrates a distributive law. The zero corresponds to a circuit that is always open, and the unit corresponds to a circuit that is always closed. The complement $\mathcal{C}'$ of a circuit $\mathcal{C}$ is open whenever $\mathcal{C}$ is closed and closed when $\mathcal{C}$ is open.

DIVISORS

As a last example, we are going to construct boolean algebras based on the divisibility relation on the set $\mathbb{P}$ of positive integers. Given two integers d and a in $\mathbb{P}$, we write $d|a$ (and call d a **divisor** of a) if $a = qd$ for some $q \in \mathbb{P}$. If $p \geqslant 2$ in $\mathbb{P}$, and the only divisors of p are 1 and p, then p is called a **prime**. Thus, the first few primes are $2, 3, 5, 7, 11, \ldots$. A fundamental fact about $\mathbb{P}$ is the **prime factorization theorem**: Every number $a \in \mathbb{P}$ is uniquely a product of primes.*

For example, the prime factorizations of 110 and 12 are $110 = 2 \cdot 5 \cdot 11$ and $12 = 2^2 \cdot 3$. If $a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$ is the prime factorization of $a \in \mathbb{P}$ where the p_i are distinct primes, the divisors d of a can be described as follows:

$$d|a \quad \text{if and only if} \quad d = p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r} \quad \text{where} \quad 0 \leqslant d_i \leqslant a_i \text{ for each } i.$$

Hence the divisors of $12 = 2^2 3^1$ in $\mathbb{P}$ are $1 = 2^0 3^0$, $2 = 2^1 3^0$, $4 = 2^2 3^0$, $3 = 2^0 3^1$, $6 = 2^1 3^1$, and $12 = 2^2 3^1$.

Given a and b in $\mathbb{P}$, let $p_1, p_2, \ldots, p_r$ denote the distinct primes that are divisors of either a or b. Hence we can write $a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}$ and $b = p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r}$, where $a_i \geqslant 0$ and $b_i \geqslant 0$ for each i. Then the **greatest common divisor** $d = \gcd(a, b)$ and the **least common multiple** $m = \operatorname{lcm}(a, b)$ of a and b are defined by

$$d = p_1^{\min(a,b)} p_2^{\min(a,b)} \cdots p_r^{\min(a,b)} \quad \text{and} \quad m = p_1^{\max(a,b)} p_2^{\max(a,b)} \cdots p_r^{\max(a,b)}.$$

* See Appendix 2 for a proof of the prime factorization theorem.

It follows that d is the unique integer in $\mathbb{P}$ that is a divisor of both a and b, and is a multiple of every such common divisor (hence the name). Similarly, m is the unique integer in $\mathbb{P}$ that is a multiple of both a and b, and is a divisor of every such common multiple. For example, $\gcd(2, 3) = 1$ and $\gcd(12, 28) = 4$, while $\text{lcm}(2, 3) = 6$ and $\text{lcm}(12, 28) = 84$.

With this background, we can describe some new examples of boolean algebras. Given $n \in \mathbb{P}$, let

$$\mathbb{D}_n = \{d \in \mathbb{P} | d \text{ divides } n\}.$$

It is clear that gcd and lcm are commutative binary operations on $\mathbb{D}_n$, and it is easy to verify that the zero is 1 and the unit is n. To prove the distributive laws, let a, b, and c be elements of $\mathbb{D}_n$, and write

$$a = p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r}, \quad b = p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r}, \quad \text{and} \quad c = p_1^{c_1} p_2^{c_2} \cdots p_r^{c_r},$$

where $p_1, p_2, \ldots, p_r$ are the distinct primes dividing at least one of a, b, and c, and where $a_i \geqslant 0$, $b_i \geqslant 0$, and $c_i \geqslant 0$ for each i. Then the first distributive law states that

$$\gcd(a, \text{lcm}(b, c)) = \text{lcm}(\gcd(a, b), \gcd(a, c)).$$

If we write out the prime factorization of each side in terms of the primes p_i, this holds if and only if for each i, the powers of p_i are equal on both sides, that is,

$$\min(a_i, \max(b_i, c_i)) = \max(\min(a_i, b_i), \min(a_i, c_i)).$$

To verify this, observe first that we may assume that $b_i \leqslant c_i$ (b_i and c_i can be interchanged without changing either side), and then check the three cases $a_i \leqslant b_i$, $b_i \leqslant a_i \leqslant c_i$, and $c_i \leqslant a_i$ separately. Hence the first distributive law holds; the other distributive law and the associative laws are verified similarly. Thus ($\mathbb{D}_n$, gcd, lcm) satisfies all the axioms for a boolean algebra except for the existence of a complement.

But complements need not exist in general: For example, 6 has no complement in $\mathbb{D}_{18} = \{1, 2, 3, 6, 9, 18\}$. Indeed, if 6 has a complement $6'$ in $\mathbb{D}_{18}$, then $\gcd(6, 6') = 1$, so we must have $6' = 1$. But then $\text{lcm}(6, 6') = 6$ and this is not the unit of $\mathbb{D}_{18}$. Hence 6 has no complement, so $\mathbb{D}_{18}$ is not a boolean algebra. However, all is not lost. The problem in $\mathbb{D}_{18}$ is that the prime factorization $18 = 2 \cdot 3^2$ has a repeated prime factor. An integer $n \in \mathbb{P}$ is called **square-free** if it is a product of distinct primes with none repeated (for example, every prime is square-free, as are $6 = 2 \cdot 3$, $10 = 2 \cdot 5$, $30 = 2 \cdot 3 \cdot 5$, etc.) If n is square-free, it is routine to verify that the complement of $d \in \mathbb{D}_n$ is $d' = n/d$, and we have

Example 2.15. If $n \in \mathbb{P}$ is square-free, then ($\mathbb{D}_n$, gcd, lcm, $'$) is a boolean algebra where $d' = n/d$ for each $d \in \mathbb{D}_n$.

The interpretations of the various boolean algebra terms are given in Table 2.8.

TABLE 2.8. Dictionary of Boolean Algebra Terms

Boolean Algebra	$\mathbb{P}(X)$	Switching Circuits	Propositional Logic	$\mathbb{D}_n$
$\wedge$	$\cap$	Series	And	gcd
$\vee$	$\cup$	Parallel	Or	lcm
$'$	$-$	Opposite	Not	$a' = n/a$
0	$\emptyset$	Open	Contradiction	1
1	X	Closed	Tautology	n
$=$	$=$	Equivalent circuit	Logically equivalent	$=$

POSETS AND LATTICES

Boolean algebras were derived from the algebra of sets, and there is one important relation between sets that we have neglected to generalize to boolean algebras, namely, the inclusion relation. This relation can be defined in terms of the union operation by

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

We can define a corresponding relation $\leqslant$ on any boolean algebra $(K, \wedge, \vee, ')$ using the meet operation:

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

If the boolean algebra is the algebra of subsets of X, this relation is the usual inclusion relation.

Proposition 2.16. $A \wedge B = A$ if and only if $A \vee B = B$. Hence either of the these conditions will define the relation $\leqslant$.

Proof. If $A \wedge B = A$, then it follows from the absorption law that $A \vee B = (A \wedge B) \vee B = B$. Similarly, if $A \vee B = B$, it follows that $A \wedge B = A$. □

Proposition 2.17. If A, B, and C are elements of a boolean algebra, K, the following properties of the relation $\leqslant$ hold.

(i) $A \leqslant A$. (reflexivity)
(ii) If $A \leqslant B$ and $B \leqslant A$, then $A = B$. (antisymmetry)
(iii) If $A \leqslant B$ and $B \leqslant C$, then $A \leqslant C$. (transitivity)

Proof

(i) $A \wedge A = A$ is an idempotent law.
(ii) If $A \wedge B = A$ and $B \wedge A = B$, then $A = A \wedge B = B \wedge A = B$.
(iii) If $A \wedge B = A$ and $B \wedge C = B$, then $A \wedge C = (A \wedge B) \wedge C$
$= A \wedge (B \wedge C) = A \wedge B = A$. □

TABLE 2.9. Partial Order Relation in Various Boolean Algebras

Boolean Algebra	Algebra of Subsets	Series-Parallel Switching Circuits	Propositional Logic	Divisors of a Square-Free Integer
$A \wedge B = A$	$A \cap B = A$	$A \wedge B = A$	$(A \text{ and } B) = A$	$\gcd(a, b) = a$
$A \leqslant B$	$A \subseteq B$	$A \Rightarrow B$	$A \Rightarrow B$	$a \mid b$
A is less than or equal to B	A is a subset of B	If A is closed, then B is closed	A implies B	a divides b

A relation satisfying the three properties in Proposition 2.17 is called a **partial order relation**, and a set with a partial order on it is called a **partially ordered set** or **poset** for short. The interpretation of the partial order in various boolean algebras is given in Table 2.9.

A partial order on a finite set K can be displayed conveniently in a **poset diagram** in which the elements of K are represented by small circles. Lines are drawn connecting these elements so that there is a path from A to B that is always directed upward if and only if $A \leqslant B$. Figure 2.15 illustrates the poset diagram of the boolean algebra of subsets $(\mathcal{P}(\{a, b\}), \cap, \cup, \bar{\ })$. Figure 2.16 illustrates the boolean algebra $\mathbb{D}_{110} = \{1, 2, 5, 11, 10, 22, 55, 110\}$ of positive divisors of $110 = 2 \cdot 5 \cdot 11$. The partial order relation is divisibility, so that there is an upward path from a to b if and only if a divides b.

The following proposition shows that $\leqslant$ has properties similar to those of the inclusion relation in sets.

Proposition 2.18. If A, B, C are elements of a boolean algebra $(K, \wedge, \vee, ')$, then the following relations hold:

(i) $A \wedge B \leqslant A$.
(ii) $A \leqslant A \vee B$.
(iii) $A \leqslant C$ and $B \leqslant C$ implies that $A \vee B \leqslant C$.

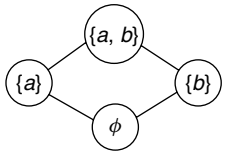


Figure 2.15. Poset diagram of $\mathcal{P}(\{a, b\})$.

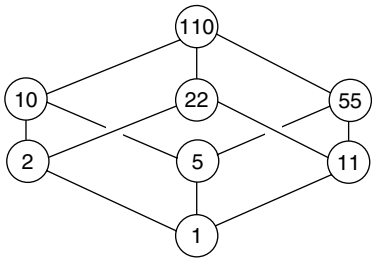


Figure 2.16. Poset diagram of $\mathbb{D}_{110}$.

(iv) $A \leqslant B$ if and only if $A \wedge B' = 0$.

(v) $0 \leqslant A$ and $A \leqslant 1$ for all A.

Proof

(i) $(A \wedge B) \wedge A = (A \wedge A) \wedge B = A \wedge B$ so $A \wedge B \leqslant A$.

(ii) $A \wedge (A \vee B) = A$, so $A \leqslant A \vee B$.

(iii) $(A \vee B) \wedge C = (A \wedge C) \vee (B \wedge C) = A \vee B$.

(iv) If $A \leqslant B$, then $A \wedge B = A$ and $A \wedge B' = A \wedge B \wedge B' = A \wedge 0 = 0$. On the other hand, if $A \wedge B' = 0$, then $A \leqslant B$ because

$$\begin{aligned} A = A \wedge 1 &= A \wedge (B \vee B') = (A \wedge B) \vee (A \wedge B') \\ &= (A \wedge B) \vee 0 = A \wedge B. \end{aligned}$$

(v) $0 \wedge A = 0$ and $A \wedge 1 = A$. □

Not all posets are derived from boolean algebras. A boolean algebra is an extremely special kind of poset. We now determine conditions which ensure that a poset is indeed a boolean algebra. Given a partial order $\leqslant$ on a set K, we have to find two binary operations that correspond to the meet and join.

An element d is said to be the **greatest lower bound** of the elements a and b in a partially ordered set if $d \leqslant a, d \leqslant b$, and x is another element, for which $x \leqslant a$, $x \leqslant b$, then $x \leqslant d$. We denote the greatest lower bound of a and b by $a \wedge b$. Similarly, we can define the **least upper bound** and denote it by $\vee$. It follows from the antisymmetry of the partial order relation that each pair of elements a and b can have at most one greatest lower bound and at most one least upper bound.

A **lattice** is a partially ordered set in which every two elements have a greatest lower bound and a least upper bound. Thus $\mathbb{D}_n$ is a lattice for every integer $n \in \mathbb{P}$, so by the discussion preceding Example 2.15, $\mathbb{D}_{18}$ is a lattice that is not a boolean algebra (see Figure 2.17).

We can now give an alternative definition of a boolean algebra in terms of a lattice: A **boolean algebra** is a lattice that has universal bounds (that is, elements 0 and 1 such that $0 \leqslant a$ and $a \leqslant 1$ for all elements a) and is distributive and complemented (that is, the distributive laws for $\wedge$ and $\vee$ hold, and complements exist). It can be verified that this definition is equivalent to our original one.

In Figure 2.18, the elements c and d have a least upper bound b but no greatest lower bound.

We note in passing that the discussion preceding Example 2.15 shows that for each $n \in \mathbb{P}$, the poset $\mathbb{D}_n$ is a lattice in which the distributive laws hold, but it is not a boolean algebra unless n is square-free. For further reading on lattices in applied algebra, consult, Davey and Priestley [16] or Lidl and Pilz [10].

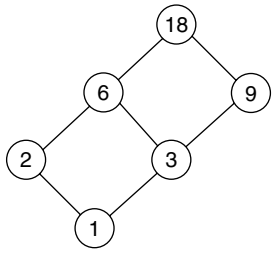


Figure 2.17. Lattice that is not a boolean algebra.

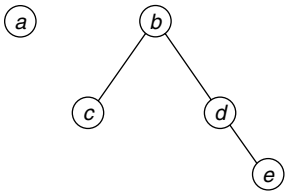


Figure 2.18. Poset that is not a lattice.

NORMAL FORMS AND SIMPLIFICATION OF CIRCUITS

If we have a complicated switching circuit represented by a boolean expression, such as

$$(A \wedge (B \vee C')') \vee ((B \wedge C') \vee A'),$$

we would like to know if we can build a simpler circuit that would perform the same function. In other words, we would like to reduce this boolean expression to a simpler form. In actual practice, it is usually desirable to reduce the circuit to the one that is cheapest to build, and the form this takes will depend on the state of the technology at the time; however, for our purposes we take the simplest form to mean the one with the fewest switches. It is difficult to find the simplest form for circuits with many switches, and there is no one method that will lead to that form. However, we do have methods for determining whether two boolean expressions are equivalent. We can reduce the expressions to a certain normal form, and the expressions will be the same if and only if their normal forms are the same. We shall look at one such form, called the **disjunctive normal form**.

In the boolean algebra of subsets of a set, every subset can be expressed as a union of singleton sets, and this union is unique to within the ordering of the terms. We shall obtain a corresponding result for arbitrary finite boolean algebras. The elements that play the role of singleton sets are called atoms. Here an **atom** in a boolean algebra $(K, \wedge, \vee, ')$ is a nonzero element B for which

$$B \wedge Y = B \quad \text{or} \quad B \wedge Y = 0 \quad \text{for each } Y \in K.$$

Thus B is an atom if $Y \leqslant B$ implies that $Y = 0$ or $Y = B$. This implies that the atoms are the elements immediately above the zero element in the poset diagram. In the case of the algebra of divisors of a square-free integer, the atoms are the primes, because the definition of b being prime is that $y|b$ implies that $y = 1$ or $y = b$.

We now give a more precise description of the algebra of switching circuits. The atoms of the algebra and the disjunctive normal form of an expression will become clear from this description.

An n-variable switching circuit can be viewed as a black box containing n independent switches $A_1, A_2, \ldots, A_n$, as shown in Figure 2.19, where each switch can be either on or off. The effect of such a circuit can be tested by trying all the 2^n different combinations of the n switches and observing when the box allows current to pass. In this way, each circuit defines a function of n variables $A_1, A_2, \ldots, A_n$:

$$f: \{0, 1\}^n \to \{0, 1\},$$

which we call the **switching function of the circuit**. Two circuits give rise to the same switching function if and only if they are equivalent.

For example, the circuit in Figure 2.20, corresponding to the expression $(A \vee B') \wedge (C \vee A')$, gives rise to the switching function $f: \{0, 1\}^3 \to \{0, 1\}$ given in Table 2.10.

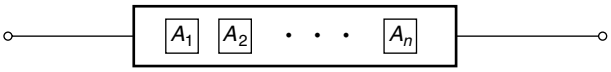


Figure 2.19. n-Variable switching circuit.

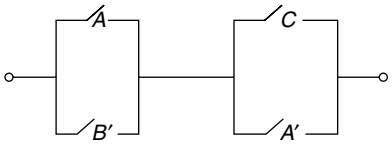


Figure 2.20. Circuit $(A \vee B') \wedge (C \vee A')$.

TABLE 2.10. Switching Function

A	B	C	$f = (A \vee B') \wedge (C \vee A')$
0	0	0	1
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

Denote the set of all n-variable switching functions from $\{0, 1\}^n$ to $\{0, 1\}$ by $\mathcal{F}_n$. Each of the 2^n elements in the domain of such a function can be mapped to either of the two elements in the codomain. Therefore, the number of different n-variable switching functions, and hence the number of different circuits with n switches, is 2^{2^n}.

Let f and g be the switching functions of two circuits of the n-variables $A_1, A_2, \ldots, A_n$. When these circuits are connected in series or in parallel, they give rise to the switching functions $f \wedge g$ or $f \vee g$, respectively, where

$$(f \wedge g)(A_1, \ldots, A_n) = f(A_1, \ldots, A_n) \wedge g(A_1, \ldots, A_n)$$

and

$$(f \vee g)(A_1, \ldots, A_n) = f(A_1, \ldots, A_n) \vee g(A_1, \ldots, A_n).$$

The switching function of the opposite circuit to that defining f is f', where

$$f'(A_1, \ldots, A_n) = (f(A_1, \ldots, A_n))'.$$

Theorem 2.19. The set of n-variable switching functions forms a boolean algebra $(\mathcal{F}_n, \wedge, \vee, ')$ that contains 2^{2^n} elements.

Proof. It can be verified that $(\mathcal{F}_n, \wedge, \vee, ')$ satisfies all the axioms of a boolean algebra. The zero element is the function whose image is always 0, and the unit element is the function whose image is always 1. □

The boolean algebra of switching functions of two variables contains 16 elements, which are displayed in Table 2.11. For example, $f_6(A, B) = 0$ if $A = B$, and 1 if $A \neq B$. This function is the exclusive OR function or a modulo 2 adder. It is also the symmetric difference function, where the symmetric difference of A and B in a boolean algebra is defined by

$$A \Delta B = (A \wedge B') \vee (A' \wedge B).$$

The operations NAND and NOR stand for "not and" and "not or," respectively; these are discussed further in the section "Transistor Gates."

As an example of the operations in the boolean algebra $\mathcal{F}_2$, we calculate the meet and join of f_{10} and f_7, and the complement of f_{10} in Table 2.12. We see that $f_{10} \wedge f_7 = f_2$, $f_{10} \vee f_7 = f_{15}$ and $f_{10}' = f_5$. These correspond to the relations $B' \wedge (A \vee B) = A \wedge B'$, $B' \vee (A \vee B) = 1$, and $(B')' = B$.

In the boolean algebra $\mathcal{F}_n$, $f \leqslant g$ if and only if $f \wedge g = f$, which happens if $g(A_1, \ldots, A_n) = 1$ whenever $f(A_1, \ldots, A_n) = 1$. The atoms of $\mathcal{F}_n$ are therefore the functions whose image contains precisely one nonzero element. $\mathcal{F}_n$ contains 2^n atoms, and the expressions that realize these atoms are of the form $A_1^{\alpha_1} \wedge A_2^{\alpha_2} \wedge \cdots \wedge A_n^{\alpha_n}$, where each $A_i^{\alpha_i} = A_i$ or A_i'.

TABLE 2.11. Two-Variable Switching Functions

A	0	0	1	1	Expressions in A and B
B	0	1	0	1	Representing the Function
f_0	0	0	0	0	0
f_1	0	0	0	1	$A \wedge B$
f_2	0	0	1	0	$A \wedge B'$ *or* $A \not\Rightarrow B$
f_3	0	0	1	1	A
f_4	0	1	0	0	$A' \wedge B$ *or* $A \not\Leftarrow B$
f_5	0	1	0	1	B
f_6	0	1	1	0	$A \Delta B$ *or* Exclusive OR(A, B)
f_7	0	1	1	1	$A \vee B$
f_8	1	0	0	0	$A' \wedge B'$ *or* NOR(A, B)
f_9	1	0	0	1	$A \Delta B'$ *or* $A \Leftrightarrow B$
f_{10}	1	0	1	0	B'
f_{11}	1	0	1	1	$A \vee B'$ *or* $A \Leftarrow B$
f_{12}	1	1	0	0	A'
f_{13}	1	1	0	1	$A' \vee B$ *or* $A \Rightarrow B$
f_{14}	1	1	1	0	$A' \vee B'$ *or* NAND(A, B)
f_{15}	1	1	1	1	1

TABLE 2.12. Some Operations in $\mathcal{F}_2$

A	B	f_{10}	f_7	$f_{10} \wedge f_7$	$f_{10} \vee f_7$	f'_{10}
0	0	1	0	0	1	0
0	1	0	1	0	1	1
1	0	1	1	1	1	0
1	1	0	1	0	1	1

The 16 elements of $\mathcal{F}_2$ are illustrated in Figure 2.21, and the four atoms are f_1, f_2, f_4, and f_8, which are defined in Table 2.11.

To show that every element of a finite boolean algebra can be written as a join of atoms, we need three preliminary lemmas.

Lemma 2.20. If $A, B_1, \ldots, B_r$ are atoms in a boolean algebra, then $A \leqslant (B_1 \vee \cdots \vee B_r)$ if and only if $A = B_i$, for some i with $1 \leqslant i \leqslant r$.

Proof. If $A \leqslant (B_1 \vee \cdots \vee B_r)$, then $A \wedge (B_1 \vee \cdots \vee B_r) = A$; thus $(A \wedge B_1) \vee \cdots \vee (A \wedge B_r) = A$. Since each B_i is an atom, $A \wedge B_i = B_i$ or 0. Not all the elements $A \wedge B_i$ can be 0, for this would imply that $A = 0$. Hence there is some i, with $1 \leqslant i \leqslant r$, for which $A \wedge B_i = B_i$. But A is also an atom, so $A = A \wedge B_i = B_i$.

The implication the other way is straightforward □

Lemma 2.21. If Z is a nonzero element of a finite boolean algebra, there exists an atom B with $B \leqslant Z$.

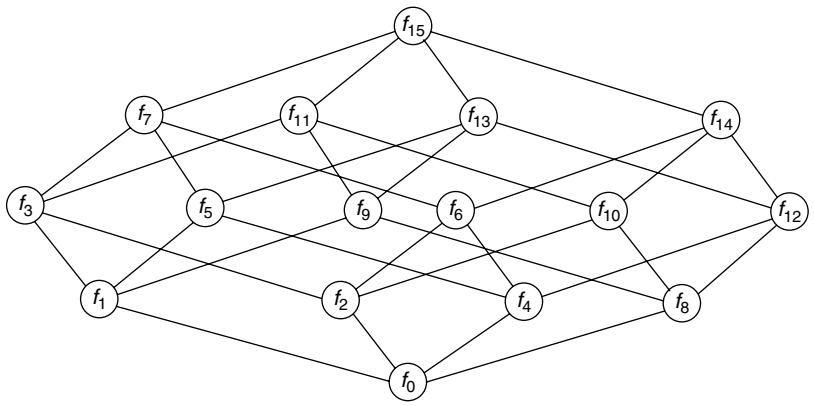


Figure 2.21. Poset diagram of the boolean algebra of two-variable switching functions.

Proof. If Z is an atom, take $B = Z$. If not, then it follows from the definition of atoms that there exists a nonzero element Z_1, different from Z, with $Z_1 \leqslant Z$. If Z_1 is not an atom, we continue in this way to obtain a sequence of distinct nonzero elements $\cdots \leqslant Z_3 \leqslant Z_2 \leqslant Z_1 \leqslant Z$, which, because the algebra is finite, must terminate in an atom B. □

Lemma 2.22. If $B_1, \ldots, B_n$ are all the atoms of a finite boolean algebra, then $Y = 0$ if and only if $Y \wedge B_i = 0$ for all i such that $1 \leqslant i \leqslant n$.

Proof. Suppose that $Y \wedge B_i = 0$ for each i. If Y is nonzero, it follows from the previous lemma that there is an atom B_j with $B_j \leqslant Y$. Hence $B_j = Y \wedge B_j = 0$, which is a contradiction, so $Y = 0$. The converse implication is trivial. □

Theorem 2.23. Disjunctive Normal Form. Each element X of a finite boolean algebra can be written as a join of atoms

$$X = B_\alpha \vee B_\beta \vee \cdots \vee B_\omega.$$

Moreover, this expression is unique up to the order of the atoms.

Proof. Let $B_\alpha, B_\beta, \ldots, B_\omega$ be all the atoms less than or equal to X in the partial order. It follows from Proposition 2.18(iii) that the join $Y = B_\alpha \vee B_\beta \vee \cdots \vee B_\omega \leqslant X$.

We will show that $X \wedge Y' = 0$, which, by Proposition 2.18(iv), is equivalent to $X \leqslant Y$. We have

$$X \wedge Y' = X \wedge B'_\alpha \wedge \cdots \wedge B'_\omega.$$

If B is an atom in the join Y, say $B = B_\alpha$, it follows that $X \wedge Y' \wedge B = 0$, since $B'_\alpha \wedge B_\alpha = 0$. If B is an atom that is not in Y, then $X \wedge Y' \wedge B = 0$ also,

because $X \wedge B = 0$. Therefore, by Lemma 2.22, $X \wedge Y' = 0$, which is equivalent to $X \leqslant Y$. The antisymmetry of the partial order relation implies that $X = Y$.

To show uniqueness, suppose that X can be written as the join of two sets of atoms

$$X = B_\alpha \vee \cdots \vee B_\omega = B_a \vee \cdots \vee B_z.$$

Now $B_\alpha \leqslant X$; thus, by Lemma 2.20, B_α is equal to one of the atoms on the right-hand side, $B_a, \ldots, B_z$. Repeating this argument, we see that the two sets of atoms are the same, except possibly for their order. □

In the boolean algebra of n-variable switching functions, the atoms are realized by expressions of the form $A_1^{\alpha_1} \wedge A_2^{\alpha_2} \wedge \cdots \wedge {A_n}^{\alpha_n}$, where the α_i's are 0 or 1 and ${A_i}^{\alpha_i} = A_i$, if $\alpha_i = 1$, whereas ${A_i}^{\alpha_i} = A_i'$, if $\alpha_i = 0$. The expression $A_1^{\alpha_1} \wedge A_2^{\alpha_2} \wedge \cdots \wedge {A_n}^{\alpha_n}$ is included in the disjunctive normal form of the function f if and only if $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = 1$. Hence there is one atom in the disjunctive normal form for each time the element 1 occurs in the image of the switching function.

Example 2.24. Find the disjunctive normal form for the expression $(B \vee (A \wedge C)) \wedge ((A \vee C) \wedge B)'$, and check the result by using the axioms to reduce the expression to that form.

Solution. We see from the values of the switching function in Table 2.13 that the disjunctive normal form is $(A' \wedge B \wedge C') \vee (A \wedge B' \wedge C)$.

From the axioms, we have

$$\begin{aligned}
(B \vee (A \wedge C)) \wedge ((A \vee C) \wedge B)' &= (B \vee (A \wedge C)) \wedge ((A' \wedge C') \vee B') \\
&= ((B \vee (A \wedge C)) \wedge (A' \wedge C')) \vee \\
&\quad ((B \vee (A \wedge C)) \wedge B') \\
&= (B \wedge A' \wedge C') \vee (A \wedge C \wedge A' \wedge C') \vee \\
&\quad (B \wedge B') \vee (A \wedge C \wedge B') \\
&= (A' \wedge B \wedge C') \vee 0 \vee 0 \vee (A \wedge B' \wedge C) \\
&= (A' \wedge B \wedge C') \vee (A \wedge B' \wedge C). \qquad \square
\end{aligned}$$

TABLE 2.13. Switching Function

A	B	C	$(B \vee (A \wedge C)) \wedge ((A \vee C) \wedge B)'$
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	0

TABLE 2.14. Switching Function

A	B	$(A \vee B) \wedge B'$	$(A \vee B) \wedge (A \wedge B)'$	$(A \wedge B)' \wedge (A \wedge B')$
0	0	0	0	0
0	1	0	1	0
1	0	1	1	1
1	1	0	0	0

Example 2.25. Determine whether any of the three expressions $(A \vee B) \wedge B'$, $(A \vee B) \wedge (A \wedge B)'$, and $(A \wedge B)' \wedge (A \wedge B')$ equivalent.

Solution. We see from Table 2.14 that $(A \vee B) \wedge B' = (A \wedge B)' \wedge (A \wedge B')$ and that these are both equal to $A \wedge B'$. □

The atoms in the boolean algebra $\mathcal{F}_2$ are realized by the expressions $A' \wedge B'$, $A' \wedge B$, $A \wedge B'$, and $A \wedge B$. These atoms partition the Venn diagram in Figure 2.22 into four disjoint regions. The disjunctive normal form for any boolean expression involving the variables A and B can be calculated by shading the region of the Venn diagram corresponding to the expression and then taking the join of the atoms in the shaded region. Figure 2.23 illustrates the eight regions of the corresponding Venn diagram for three variables.

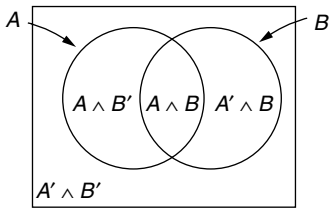


Figure 2.22. Venn diagram for $\mathcal{F}_2$.

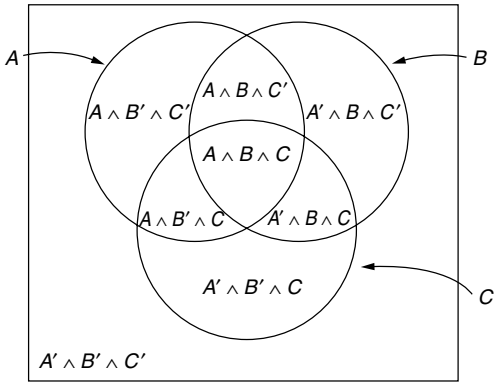


Figure 2.23. Venn diagram for $\mathcal{F}_3$.

By looking at the shaded region of a Venn diagram corresponding to a boolean expression, it is often possible to see how to simplify the expression. Furthermore, the disjunctive normal form provides a method of proving hypotheses derived from these Venn diagrams.

However, Venn diagrams become too complicated and impractical for functions of more than four variables. For other general methods of simplifying circuits, consult a book on boolean algebras such as Mendelson [21].

Example 2.26. Find the disjunctive normal form and simplify the circuit in Figure 2.24.

Solution. This circuit is represented by the boolean expression

$$f = A \vee ((B' \vee C) \wedge (A \vee B \vee C)).$$

The boolean function $f: \{0, 1\}^3 \to \{0, 1\}$ that this expression defines is given in Table 2.15. It follows that the disjunctive normal form is

$$(A' \wedge B' \wedge C) \vee (A' \wedge B \wedge C) \vee (A \wedge B' \wedge C') \vee (A \wedge B' \wedge C)$$
$$\vee (A \wedge B \wedge C') \vee (A \wedge B \wedge C),$$

which is certainly not simpler than the original. However, by looking at the Venn diagram in Figure 2.25, we see that this expression is equivalent to just $A \vee C$; thus a simpler equivalent circuit is given in Figure 2.25. □

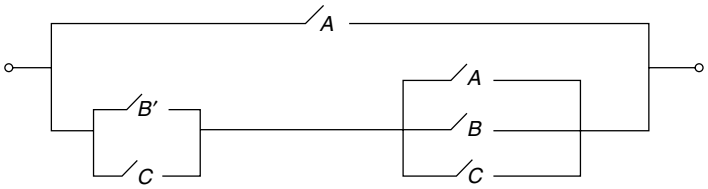


Figure 2.24. Series-parallel circuit.

TABLE 2.15. Switching Function

A	B	C	f
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

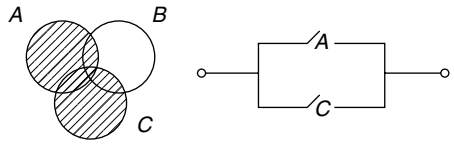


Figure 2.25. Venn diagram and simplified circuit.

In building a computer, one of the most important pieces of equipment needed is a circuit that will add two numbers in binary form. Consider the problem of adding the numbers 15 and 5. Their binary forms are 1111 and 101, respectively. The binary and decimal additions are shown below. In general, if we add the number $\ldots a_2a_1a_0$ to $\ldots b_2b_1b_0$, we have to carry the digits $\ldots c_2c_1$ to obtain the sum $\ldots s_2s_1s_0$.

1111		15	$\ldots a_2a_1a_0$
101		5	$\ldots b_2b_1b_0$
1111	$\leftarrow$ carry digits $\rightarrow$	1	$\ldots c_2c_1$
10100		20	$\ldots s_2s_1s_0$
binary addition		decimal addition	

Let us first design a circuit to add a_0 and b_0 to obtain s_0 and the carry digit c_1. This is called a **half adder**. The digits s_0 and c_1 are functions of a_0 and b_0 which are given by Table 2.16. For example, in binary arithmetic, $1 + 1 = 10$, which means that if $a_0 = 1$ and $b_0 = 1$, $s_0 = 0$, and we have to carry $c_1 = 1$.

We see from Table 2.16 that $c_1 = a_0 \wedge b_0$ and $s_0 = (a_0' \wedge b_0) \vee (a_0 \wedge b_0')$. These circuits are shown in Figure 2.26.

TABLE 2.16. Switching Functions for the Half Adder

a_0	b_0	c_1	s_0
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

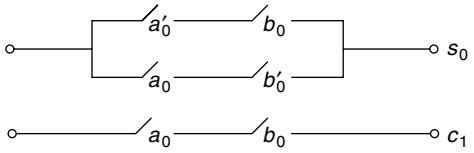


Figure 2.26. Circuits for the half adder.

TABLE 2.17. Switching Functions for a Full Adder

a_i	b_i	c_i	c_{i+1}	s_i
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

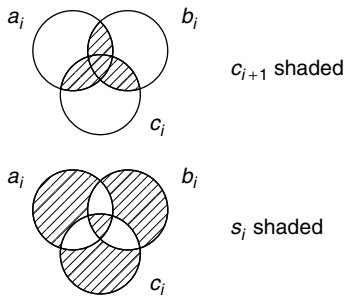


Figure 2.27. Venn diagrams for a full adder.

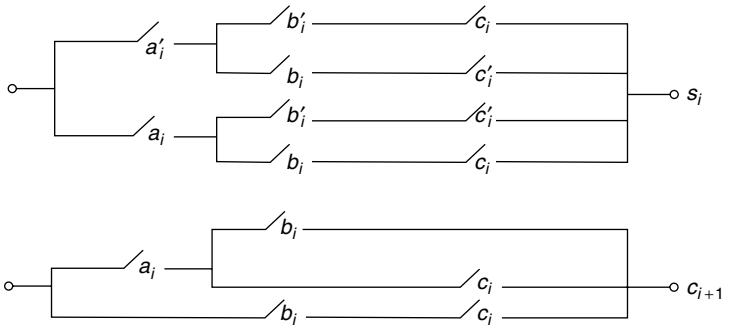


Figure 2.28. Circuits for a full adder.

A circuit that adds a_i, b_i, and the carry digit, c_i, to obtain s_i, with c_{i+1} to carry, is called a **full adder**. The functions c_{i+1} and s_i are defined by Table 2.17, and their Venn diagrams are given in Figure 2.27. Notice that $s_i = a_i \Delta b_i \Delta c_i$.

Suitable expressions for a full adder are as follows. The corresponding circuits are shown in Figure 2.28.

$$
\begin{aligned}
s_i &= (a_i' \wedge b_i' \wedge c_i) \vee (a_i' \wedge b_i \wedge c_i') \vee (a_i \wedge b_i' \wedge c_i') \vee (a_i \wedge b_i \wedge c_i) \\
&= (a_i' \wedge ((b_i' \wedge c_i) \vee (b_i \wedge c_i'))) \vee (a_i \wedge ((b_i' \wedge c_i') \vee (b_i \wedge c_i))). \\
c_{i+1} &= (a_i \wedge b_i) \vee (a_i \wedge c_i) \vee (b_i \wedge c_i) \\
&= (a_i \wedge (b_i \vee c_i)) \vee (b_i \wedge c_i).
\end{aligned}
$$

Using one half adder and $(n - 1)$ full adders, we can design a circuit that will add two numbers that in binary form have n or fewer digits (that is, numbers less than 2^n).

TRANSISTOR GATES

The switches we have been dealing with so far have been simple two-state devices. Transistor technology, however, allows us to construct basic switches with multiple inputs. These are called *transistor gates*. Transistor gates can be used to implement the logical operations AND, OR, NOT, and modulo 2 addition (that is, exclusive OR). Gates for the composite operations NOT-AND and NOT-OR are also easily built from transistors; these are called NAND and NOR gates, respectively. Figure 2.29 illustrates the symbols and outputs for these gates when there are two inputs. However, any number of inputs is possible. Note that the inversion operation is indicated by a small circle.

Transistor gates can be combined in series and in parallel to form more complex circuits. Any circuit with n inputs and one output defines an n-variable switching function. The set of all such n-variable functions again forms the boolean algebra $(\mathcal{F}_n, \wedge, \vee, ')$.

It follows from the disjunctive normal form that any boolean function can be constructed from AND, OR, and NOT gates. What is not so obvious is that any boolean function can be constructed solely from NOR gates (or solely from NAND gates). This is of interest because with certain types of transistors, it is easier to build NOR gates (and NAND gates) than it is to build the basic operations. Figure 2.30 illustrates how two-input NOR gates can be used to construct two-input AND and OR gates as well as a NOT gate.

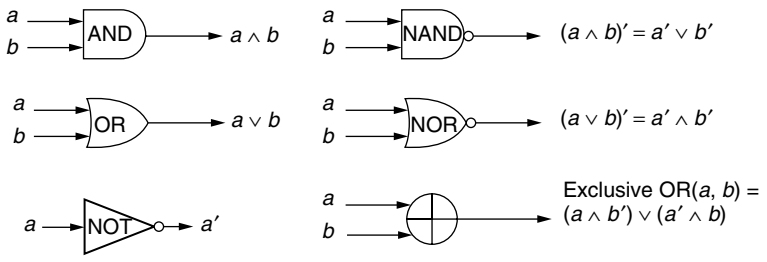


Figure 2.29. Transistor gates.

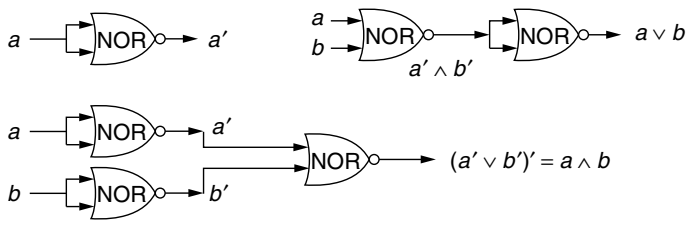


Figure 2.30. Basic operations constructed from NOR gates.

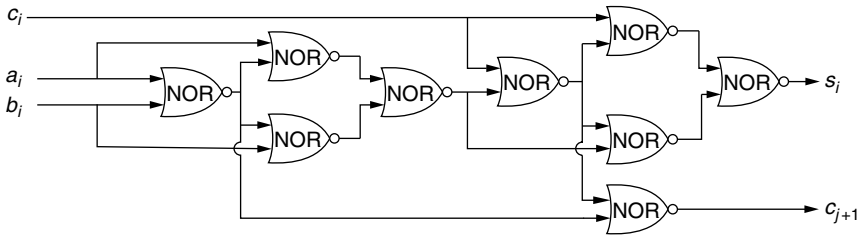


Figure 2.31. Full adder using NOR gates.

Example 2.27. Verify that the circuit in Figure 2.31 is indeed a full adder.

Solution. We analyze this circuit by breaking it up into component parts as illustrated in Figure 2.32. Consider the subcircuit consisting of four NOR gates in Figure 2.32 with inputs a and b and outputs l and m. If u and v are the intermediate functions as shown in the figure, then

$$\begin{aligned} m &= \mathrm{NOR}(a, b) = a' \wedge b' \\ u &= \mathrm{NOR}(a, a' \wedge b') = a' \wedge (a' \wedge b')' = a' \wedge (a \vee b) \\ &= (a' \wedge a) \vee (a' \wedge b) = 0 \vee (a' \wedge b) = a' \wedge b, \end{aligned}$$

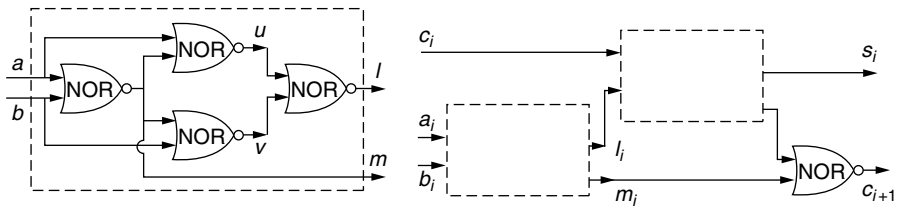


Figure 2.32. Component parts of the full adder.

and $v = a \wedge b'$, similarly. Therefore,

$$l = \text{NOR}(u, v) = (a' \wedge b)' \wedge (a \wedge b')' = (a \vee b') \wedge (a' \vee b)$$
$$= (a \wedge a') \vee (a \wedge b) \vee (b' \wedge a') \vee (b' \wedge b) = 0 \vee (a \wedge b) \vee (b' \wedge a') \vee 0$$
$$= (a \wedge b) \vee (a' \wedge b') = a \Delta b'.$$

The entire circuit can now be constructed from two of these identical subcircuits together with one NOR gate, as shown in Figure 2.32. The switching functions for the subcircuit and the full adder are calculated in Table 2.18.

We have $c_{i+1} = \text{NOR}(m_i, \text{NOR}(c_i, l_i))$, while $s_i = c_i \Delta l_i'$. We see from Table 2.18 that the circuits do perform the addition of a_i, b_i, and c_i correctly. □

TABLE 2.18. Switching Functions for the NOR Circuit

a	b	l	m
0	0	1	1
0	1	0	0
1	0	0	0
1	1	1	0

a_i	b_i	c_i	l_i	m_i	$\text{NOR}(c_i, l_i)$	c_{i+1}	s_i
0	0	0	1	1	0	0	0
0	0	1	1	1	0	0	1
0	1	0	0	0	1	0	1
0	1	1	0	0	0	1	0
1	0	0	0	0	1	0	1
1	0	1	0	0	0	1	0
1	1	0	1	0	0	1	0
1	1	1	1	0	0	1	1

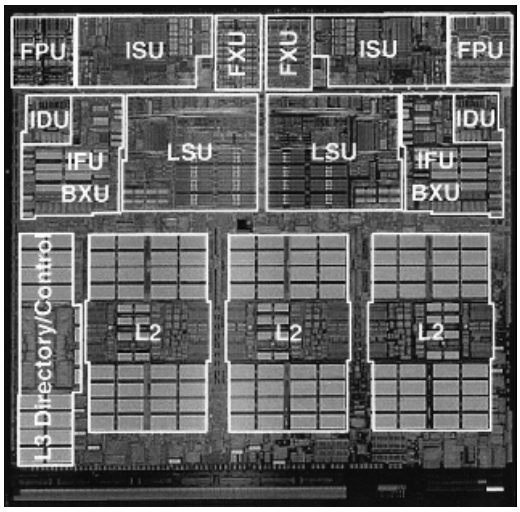


Figure 2.33. Photomicrograph of the IBM POWER4 chip containing 174 million transistors. (Courtesy of the *IBM Journal of Research and Development*.)

Instead of using many individual transistors, circuits are now made on a single semiconductor "chip," such as the one in Figure 2.33. This chip may contain millions of gates and several layers of semiconductor. Simplification of a circuit may not mean the reduction of the circuit to the smallest number of gates. It could mean simplification to standard modules, or the reduction of the numbers of layers in the chip. In the design of high-speed computers, it is important to reduce the time a circuit will take to perform a given set of operations.

REPRESENTATION THEOREM

A boolean algebra is a generalization of the notion of the algebra of sets. However, we now show that every *finite* boolean algebra is in fact essentially the same as the algebra of subsets of some finite set. To be more precise about what we mean by algebras being essentially the same, we introduce the notion of morphism and isomorphism of boolean algebras. A morphism between two boolean algebras is a function between their elements that preserves the two binary operations and the unary operation.

More precisely, if $(K, \wedge, \vee, ')$ and $(L, \cap, \cup, ^-)$ are two boolean algebras, the function $f: K \to L$ is called a **boolean algebra morphism** if the following conditions hold for all $A, B \in K$:

(i) $f(A \wedge B) = f(A) \cap f(B)$.
(ii) $f(A \vee B) = f(A) \cup f(B)$.
(iii) $f(A') = \overline{f(A)}$.

A boolean algebra **isomorphism** is a bijective boolean algebra morphism.

Isomorphic boolean algebras have identical properties. For example, their poset diagrams are the same, except for the labeling of the elements. Furthermore, the atoms of one algebra must correspond to the atoms in the isomorphic algebra.

If we wish to find an isomorphism between any boolean algebra, K, and an algebra of sets, the atoms of K must correspond to the singleton elements of the algebra of sets. This suggests that we try to define an isomorphism from K to the algebra $\mathcal{P}(A)$ of subsets of the set A of atoms of K. The following theorem shows that if K is finite, we can set up such an isomorphism.

Theorem 2.28. Representation Theorem for Finite Boolean Algebras. Let A be the set of atoms of the finite boolean algebra $(K, \wedge, \vee, ')$. Then there is a boolean algebra isomorphism between $(K, \wedge, \vee, ')$ and the algebra of subsets $(\mathcal{P}(A), \cap, \cup, ^-)$.

Proof. We already have a natural correspondence between the atoms of K and the atoms of $\mathcal{P}(A)$. We use the disjunctive normal form (Theorem 2.23) to extend this correspondence to all the elements of K.

By the disjunctive normal form, any element of K can be written as a join of atoms of K, say $B_\alpha \vee \cdots \vee B_\omega$. Define the function $f: K \to \mathcal{P}(A)$ by

$$f(B_\alpha \vee \cdots \vee B_\omega) = \{B_\alpha\} \cup \cdots \cup \{B_\omega\}.$$

The uniqueness of the normal form implies that each element of K has a unique image in $\mathcal{P}(A)$ and that f is a bijection.

We still have to show that f is a morphism of boolean algebras. If X and Y are two elements of K, the atoms in the normal forms of $X \vee Y$ and $X \wedge Y$ are, respectively, the atoms in the forms of X or Y and the atoms common to the forms of X and Y. Therefore, $f(X \vee Y) = f(X) \cup f(Y)$, and $f(X \wedge Y) = f(X) \cap f(Y)$. An atom B is in the normal form for X' if and only if $B \leqslant X'$, which, by Proposition 2.18(iv), happens if and only if $B \wedge X = 0$. Therefore, the atoms in X' are all the atoms that are not in X and $f(X') = \overline{f(X)}$. This proves that f is a boolean algebra isomorphism. □

COROLLARY 2.29. If $(K, \wedge, \vee, ')$ is a finite boolean algebra, then K has 2^n elements, where n is the number of atoms in K.

Proof. This follows from Theorem 2.5. □

Consider the representation theorem (Theorem 2.28) applied to the boolean algebra $\mathbb{D}_{110}$, which consists of the divisors of 110. The atoms of this algebra are the prime divisors 2, 5, and 11. Theorem 2.28 defines a boolean algebra isomorphism to the algebra of subsets of $\{2, 5, 11\}$. This isomorphism, f, maps a number onto the subset consisting of its prime divisors; for example, $f(11) = \{11\}$ and $f(10) = \{2, 5\}$.

Example 2.30. Do the divisors of 12 form a boolean algebra under gcd and lcm?

Solution. The set of divisors of 12 is $\{1, 2, 3, 4, 6, 12\}$. Since the number of elements is not a power of 2, it cannot form a boolean algebra. □

Example 2.31. Do the divisors of 24 form a boolean algebra under gcd and lcm?

Solution. There are $8 = 2^3$ divisors of 24, namely 1, 2, 3, 4, 6, 8, 12, and 24. However, the poset diagram in Figure 2.34 shows that 2 and 3 are the only atoms. Hence, by Corollary 2.29, it cannot be a boolean algebra because it does not have $2^2 = 4$ elements. □

An *infinite* boolean algebra is not necessarily isomorphic to the algebra of *all* subsets of a set, but is isomorphic to the algebra of *some* subsets of a set. This result is known as *Stone's representation theorem*, and a proof can be found in Mendelson [21, Sec. 5.7].

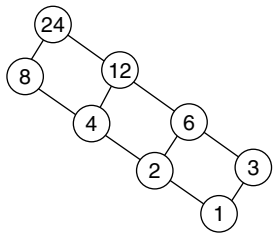


Figure 2.34. Poset diagram of the divisors of 24.

EXERCISES

If A, B, and C are subsets of a set X, under what conditions do the equalities in Exercises 2.1 to 2.6 hold?

2.1. $A \cup B = A \Delta B \Delta (A \cap B)$.

2.2. $A \cap (B \cup C) = (A \cap B) \cup C$.

2.3. $A - (B \cup C) = (A - B) \cup (A - C)$.

2.4. $A \Delta (B \cap C) = (A \Delta B) \cap (A \Delta C)$.

2.5. $A \Delta (B \cup C) = (A \Delta B) \cup (A \Delta C)$.

2.6. $A \cup (B \cap A) = A$.

2.7. Prove the remaining parts of Proposition 2.1.

2.8. Prove the remaining parts of Proposition 2.3.

2.9. Prove Theorem 2.5 by induction on n. That is, if X is a finite set with n elements, prove that $\mathcal{P}(X)$ contains 2^n elements.

2.10. Prove or give a counterexample to the following statements.
(a) $\mathcal{P}(X) \cap \mathcal{P}(Y) = \mathcal{P}(X \cap Y)$.
(b) $\mathcal{P}(X) \cup \mathcal{P}(Y) = \mathcal{P}(X \cup Y)$.

2.11. (**Cantor's theorem**) Prove that there is no surjective (onto) function from X to $\mathcal{P}(X)$ for any finite or infinite set X. This shows that $\mathcal{P}(X)$ always contains more elements than X.
[*Hint*: If $f: X \to \mathcal{P}(X)$, consider $\{x \in X | x \notin f(x)\}$.]

2.12. Write down the table for $(\mathcal{P}(X), -)$, under the difference operation, when $X = \{a, b\}$.

2.13. If A, B, C, and D are finite sets, find an expression for $|A \cup B \cup C \cup D|$ in terms of the number of elements in their intersections.

2.14. Of the Chelsea pensioners who returned from a war, *at least* 70% had lost an eye, 75% an ear, 80% an arm, and 85% a leg. What percentage, *at least* must have lost all four? (From Lewis Carroll, *A Tangled Tale*.)

2.15. One hundred students were questioned about their study habits. Seventy said they sometimes studied during the day, 55 said they sometimes studied during the night, and 45 said they sometimes studied during the weekend.

Also, 36 studied during the day and night, 24 during the day and at weekends, 17 during the night and at weekends, and 3 during the day, night, and weekends. How many did not study at all?

2.16. Prove that the associative laws in defined in the section "Boolean Algebras" follow from the other axioms of a boolean algebra.

2.17. If the zero element is the same as the unit element in a boolean algebra, prove that the algebra has only one element. Is this algebra isomorphic to the algebra of subsets of some set?

2.18. Draw the poset diagram for $\mathcal{F}_1$, the boolean algebra of switching functions of one variable.

If A, B, and C are elements of a boolean algebra $(K, \wedge, \vee, ')$ *and* $\leqslant$ *is the related partial order, prove the assertions in Exercises 2.19 to 2.24 from the axioms and Propositions 2.13, 2.17, and 2.18.*

2.19. $0' = 1$.

2.20. $A \wedge (A' \vee B) = A \wedge B$.

2.21. $(A \wedge B) \vee (B \wedge C) \vee (C \wedge A) = (A \vee B) \wedge (B \vee C) \wedge (C \vee A)$.

2.22. $A \leqslant B \wedge C$ implies that $A \leqslant B$ and $A \leqslant C$.

2.23. $(A \wedge B') \vee C = (B \wedge C) \vee (B' \wedge (A \vee C))$.

2.24. $A \leqslant B$ if and only if $B' \leqslant A'$.

2.25. Write down the truth tables for the following propositions. Which of these propositions are equivalent?

(a) $A \Rightarrow B$. **(b)** $B' \Rightarrow A'$.
(c) $(A \wedge B) \Leftrightarrow B$. **(d)** $(A \vee B) \Leftrightarrow A$.

2.26. Is the proposition $[(A' \Rightarrow B) \wedge (B \Leftrightarrow C)']$ equivalent to $[B \vee (A \wedge C)]$ or $[(C \Rightarrow B) \vee (B \Rightarrow (A \wedge C))]$?

2.27. Which of the following are tautologies, and which are contradictions?

(a) $(A \wedge B) \Leftrightarrow (A \Rightarrow B')$. **(b)** $A \Rightarrow (B \Rightarrow A)$.
(c) $(A \wedge B') \Leftrightarrow (A \Rightarrow B)'$. **(d)** $(A \Rightarrow B) \Rightarrow ((B \Rightarrow C) \Rightarrow (A \Rightarrow C))$.

2.28. Harry broke the window if and only if he ran away and John was lying. John said that either Harry broke the window or Harry did not run away. If Harry ran away, then he did not break the window. What conclusions can you come to?

Draw circuits to realize the expressions in Exercises 2.29 and 2.30.

2.29. $(A \wedge (B \vee C' \vee D)) \vee B'$.

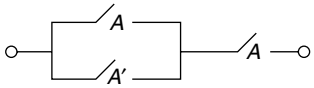
2.30. $(A \wedge B' \wedge C') \vee (A' \wedge B \wedge C) \vee (A \wedge B \wedge C)$.

2.31. Simplify the following expression and then draw a circuit for it.

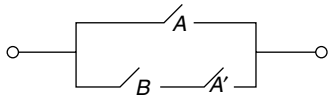
$$((A \wedge B) \vee C') \wedge (B' \vee (C \wedge A')) \vee (A' \wedge B' \wedge C').$$

Give a boolean expression for each of the circuits in Exercises 2.32 to 2.36, find their disjunctive normal forms, and then try to simplify the circuits.

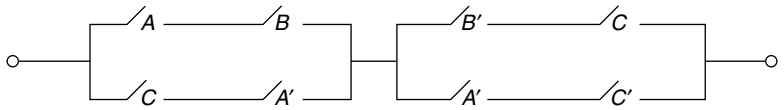
2.32.



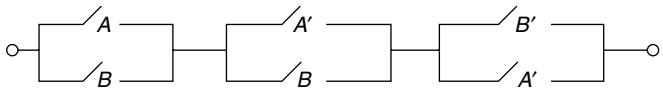
2.33.



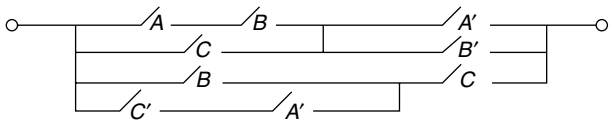
2.34.



2.35.



2.36.



2.37. By looking at all the possible paths through the **bridge circuit** in Figure 2.35, show that it corresponds to the boolean expression

$$(A \wedge D) \vee (B \wedge E) \vee (A \wedge C \wedge E) \vee (B \wedge C \wedge D).$$

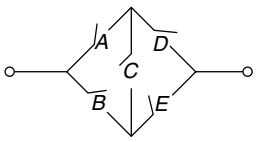


Figure 2.35

2.38. Find a series-parallel circuit that is equivalent to the bridge circuit in Figure 2.36 and simplify your circuit.

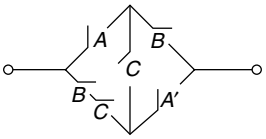


Figure 2.36

2.39. A hall light is controlled by two switches, one upstairs and one downstairs. Design a circuit so that the light can be switched on or off from the upstairs or the downstairs.

2.40. A large room has three separate entrances, and there is a light switch by each entrance. Design a circuit that will allow the lights to be turned on or off by throwing any one switch.

2.41. A voting machine for three people contains three YES–NO switches and allows current to pass if and only if there is a majority of YES votes. Design and simplify such a machine.

2.42. Design and simplify a voting machine for five people.

2.43. Design a circuit for a light that is controlled by two independent switches A and B and a master switch C. C must always be able to turn the light on. When C is off, the light should be able to be turned on and off using A or B.

2.44. A committee consists of a chairman A, and three other members, B, C, and D. If B, C, and D, are not unanimous in their voting, the chairman decides the vote. Design a voting machine for this committee and simplify it as much as possible.

2.45. Verify that the Venn diagram in Figure 2.37 illustrates the 16 atoms for a boolean expression in four variables. Then use the diagram to simplify the circuit in Figure 2.37.

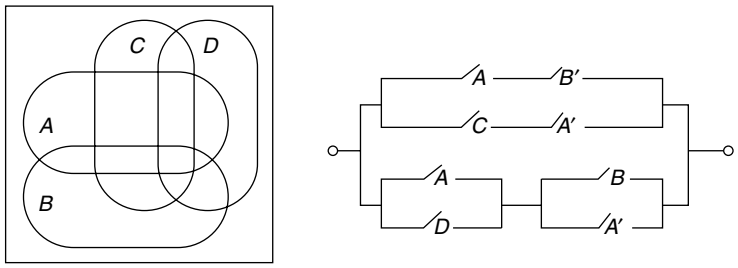


Figure 2.37

2.46. Design four series-parallel circuits to multiply two numbers in binary form that have at most two digits each.

2.47. Design a circuit that will turn an orange light on if exactly one of the four switches A, B, C, and D is on and a green light when all four are on.

2.48. Five switches are set to correspond to a number in binary form that has at most five digits. Design and simplify a circuit that will switch a light on if and only if the binary number is a perfect square.

2.49. In Chapter 11 we construct a finite field $F = \{0, 1, \alpha, \beta\}$ whose multiplication table is given in Table 2.19. Writing 00 for 0, 01 for 1, 10 for α, and 11 for β, design and simplify circuits to perform this multiplication.

2.50. A swimming pool has four relay switches that open when the water temperature is above the maximum allowable, when the water temperature is below the minimum, when the water level is too high, and when the level

TABLE 2.19. Multiplication in a Four-Element Field

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

is too low. These relays are used to control the valves that add cold water, that let water out, and that heat the water in the pool. Design and simplify a circuit that will perform the following tasks. If the temperature is correct but the level too high, it is to let water out. If the temperature is correct but the level too low, it is to let in cold water and heat the water. If the pool is too warm, add cold water and, if the level is also too high, let water out at the same time. If the pool is too cold but the level correct, heat the water; if the level is too low, heat the water and add cold water, and, if the level is too high, just let out the water.

2.51. In a dual fashion to the disjunctive normal form, every boolean expression in n-variables can be written in its *conjunctive normal form*. What are the conjunctive normal forms for $A \Delta B$ and $A \wedge B'$?

Draw poset diagrams for the sets given in Exercises 2.52 to 2.57 with divisibility as the partial order and determine whether the systems are lattices or boolean algebras.

2.52. $\{1, 2, 3, 4, 5, 6\}$.

2.53. $\{2, 4, 6, 12\}$.

2.54. $\mathbb{D}_{54}$

2.55. $\{1, 2, 4, 8\}$.

2.56. $\mathbb{D}_{42}$

2.57. $\{1, 2, 3, 5, 6, 10, 30, 60\}$.

2.58. Prove that $(\mathbb{D}_n, |)$ is a poset for each $n \geqslant 2$, and prove the distributive laws.

2.59. Let $n = p_1 p_2 \cdots p_r$, where the p_i are distinct primes. Describe the atoms in the boolean algebra $(\mathbb{D}_n, \text{gcd}, \text{lcm}, ')$.

2.60. Prove that an element $A \neq 0$ in a boolean algebra is an atom if and only if for each B in the algebra, either $A \leqslant B$ or $A \leqslant B'$.

2.61. Suppose that A and B are elements of a boolean algebra. If an element X in the algebra exists such that $A \wedge X = B \wedge X$ and $A \vee X = B \vee X$, show that $A = B$. [*Hint*: $A = A \wedge (A \vee X)$.]

2.62. Let $K = \{x \in \mathbb{R} | 0 \leqslant x \leqslant 1\}$ and let $x \wedge y$ and $x \vee y$ be the smaller and larger of x and y, respectively. Show that it is not possible to define a complement $'$ on K so that $(K, \wedge, \vee, ')$ is a boolean algebra. However, if we define $x' = 1 - x$, which of the properties defined in the section "Boolean Algebras" and in Proposition 2.13 remain true? This is the kind of algebraic model that would be required to deal with transistor switching gates under transient conditions. The voltage or current varies continuously between the levels 0 and 1, while an AND gate performs the operation $x \wedge y$, an OR gate performs $x \vee y$, and a NOT gate produces x'.

2.63. If f is a boolean algebra morphism from K to L, prove that $f(0_K) = 0_L$ and $f(1_K) = 1_L$, where $0_K, 0_L, 1_K$, and 1_L are the respective zero and unit elements.

2.64. Write down the tables for the NOR and NAND operations on the set $\mathcal{P}(\{a, b, c\})$.

2.65. Can every switching circuit be built out of AND and NOT gates?

2.66. **(a)** Design a half adder using five NOR gates.
(b) Design a half adder using five NAND gates.

2.67. Analyze the effect of the circuit in Figure 2.38.

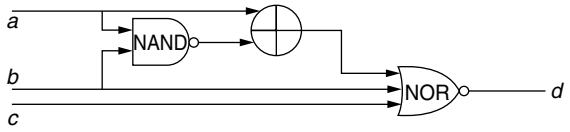


Figure 2.38

2.68. Design a NOR circuit that will produce a parity check symbol for four binary input digits; that is, the circuit must produce a 0 if the inputs contain an even number of 1's, and it must produce 1 otherwise.

2.69. One of the basic types of components of a digital computer is a **flip-flop**. This is a device that can be in one of two states (corresponding to outputs 0 and 1) and it will remain in a particular state Q until an input changes the state to the next state Q^*. One important use of a flip-flop is to store a binary digit. An RS flip-flop is a circuit with two inputs, R and S, and one output, Q, corresponding to the state of the flip-flop. An input $R = 1$ resets the next state Q^* to 0, and an input $S = 1$ sets the next state to 1. If both R and S are 0, the next state is the same as the previous state Q. It is assumed that R and S cannot both be 1 simultaneously. Verify that the NOR circuit in Figure 2.39 is indeed an RS flip-flop. To eliminate spurious effects due to the time it takes a transistor to operate, this circuit should be controlled by a "clock." The output Q should be read only when the clock "ticks," whereas the inputs are free to change between ticks.

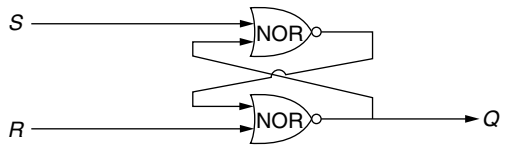


Figure 2.39. *RS* flip-flop.

2.70. A JK flip-flop is similar to an RS flip-flop except that both inputs are allowed to be 1 simultaneously, and in this case the state Q changes to its opposite state. Design a JK flip-flop using NOR and NAND gates.

3

GROUPS

Symmetries and permutations in nature and in mathematics can be described conveniently by an algebraic object called a *group*. In Chapter 5, we use group theory to determine all the symmetries that can occur in two- or three-dimensional space. This can be used, for example, to classify all the forms that chemical crystals can take. If we have a large class of objects, some of which are equivalent under permutations or symmetries, we show, in Chapter 6, how groups can be used to count the nonequivalent objects. For example, we count the number of different switching functions of n variables if we allow permutations of the inputs.

Historically, the basic ideas of group theory arose with the investigation of permutations of finite sets in the theory of equations. One of the aims of mathematicians at the beginning of the nineteenth century was to find methods for solving polynomial equations of degree 5 and higher. Algorithms, involving the elementary arithmetical operations and the extraction of roots, were already known for solving all polynomial equations of degree less than 5; the formulas for solving quadratic equations had been known since Babylonian times, and cubic and quartic equations had been solved by various Italian mathematicians in the sixteenth century. However, in 1829, using the rudiments of group theory, the Norwegian Niels Abel (1802–1829) showed that some equations of the fifth degree could not be solved by any such algorithm. Just before he was mortally wounded in a duel, at the age of 20, the brilliant mathematician Évariste Galois (1811–1832) developed an entire theory that connected the solvability of an equation with the permutation group of its roots. This theory, now called Galois theory, is beyond the scope of this book, but interested students should look at Stewart [35] after reading Chapter 11.

It was not until the 1880s that the abstract definition of a group that we use today began to emerge. However, Cayley's theorem, proved at the end of this chapter, shows that every abstract group can be considered as a group of permutations. It was soon discovered that this concept of a group was so universal that it cropped up in many different branches of mathematics and science.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

GROUPS AND SYMMETRIES

A **group** $(G, \cdot)$ is a set G together with a binary operation $\cdot$ satisfying the following axioms.

(i) G is **closed** under the operation $\cdot$; that is, $a \cdot b \in G$ for all $a, b \in G$.

(ii) The operation $\cdot$ is **associative**; that is, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ for all $a, b, c \in G$.

(iii) There is an **identity element** $e \in G$ such that $e \cdot a = a \cdot e = a$ for all $a \in G$.

(iv) Each element $a \in G$ has an **inverse element** $a^{-1} \in G$ such that $a^{-1} \cdot a = a \cdot a^{-1} = e$.

The closure axiom is already implied by the definition of a binary operation; however, it is included because it is often overlooked otherwise.

If the operation is commutative, that is, if $a \cdot b = b \cdot a$ for all $a, b \in G$, the group is called **commutative** or **abelian**, in honor of the mathematician Niels Abel.

Let G be the set of complex numbers $\{1, -1, i, -i\}$ and let $\cdot$ be the standard multiplication of complex numbers. Then $(G, \cdot)$ is an abelian group. The product of any two of these elements is an element of G; thus G is closed under the operation. Multiplication is associative and commutative in G because multiplication of complex numbers is always associative and commutative. The identity element is 1, and the inverse of each element a is the element $1/a$. Hence $1^{-1} = 1$, $(-1)^{-1} = -1$, $i^{-1} = -i$, and $(-i)^{-1} = i$. The multiplication of any two elements of G can be represented by Table 3.1.

The set of all rational numbers, $\mathbb{Q}$, forms an abelian group $(\mathbb{Q}, +)$ under addition. The identity is 0, and the inverse of each element is its negative. Similarly, $(\mathbb{Z}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$ are all abelian groups under addition.

If $\mathbb{Q}^*$, $\mathbb{R}^*$, and $\mathbb{C}^*$ denote the set of nonzero rational, real, and complex numbers, respectively, $(\mathbb{Q}^*, \cdot)$, $(\mathbb{R}^*, \cdot)$, and $(\mathbb{C}^*, \cdot)$ are all abelian groups under multiplication.

For any set X, $(\mathcal{P}(X), \Delta)$ is an abelian group. The group axioms follow from Proposition 2.3; the empty set, $\emptyset$, is the identity, and each element is its own inverse.

TABLE 3.1. Group $\{1, -1, i, -i\}$

$\cdot$	1	-1	i	$-i$
1	1	-1	i	$-i$
-1	-1	1	$-i$	i
i	i	$-i$	-1	1
$-i$	$-i$	i	1	-1

Every group must have at least one element, namely, its identity, e. A group with only this one element is called **trivial**. A trivial group takes the form $(\{e\}, \cdot)$, where $e \cdot e = e$.

Many important groups consist of functions. Given functions $f: X \to Y$ and $g: Y \to Z$, their **composite** $g \circ f: X \to Z$ is defined by

$$(g \circ f)(x) = g(f(x)) \qquad \text{for all } x \in X.$$

Composition is associative; that is, if $h: Z \to W$, then $h \circ (g \circ f) = (h \circ g) \circ f$. Indeed,

$$(h \circ (g \circ f))(x) = h(g(f(x))) = ((h \circ g) \circ f)(x)$$

for all $x \in X$, as is readily verified. In particular, if X is a set, then $\circ$ is an associative binary operation on the set of all functions $f: X \to X$. Moreover, this operation has an identity: The **identity function** $1_X: X \to X$ is defined by

$$1_X(x) = x \qquad \text{for all } x \in X.$$

Then $1_X \circ f = f = f \circ 1_X$ for all $f: X \to X$. Hence, we say that a function $f': X \to X$ is an **inverse** of $f: X \to X$ if

$$f' \circ f = 1_X \quad \text{and} \quad f \circ f' = 1_X;$$

equivalently if $f'(f(x)) = x$ and $f(f'(x)) = x$ for all $x \in X$. This inverse is unique when it exists. For if f'' is another inverse of f, then

$$f' = f' \circ 1_X = f' \circ (f \circ f'') = (f' \circ f) \circ f'' = 1_X \circ f'' = f''.$$

When it exists (see Theorem 3.3) the inverse of f is denoted f^{-1}.

Example 3.1. A translation of the plane $\mathbb{R}^2$ in the direction of the vector (a, b) is a function $f: \mathbb{R}^2 \to \mathbb{R}^2$ defined by $f(x, y) = (x + a, y + b)$. The composition of this translation with a translation g in the direction of (c, d) is the function $f \circ g: \mathbb{R}^2 \to \mathbb{R}^2$, where

$$f \circ g(x, y) = f(g(x, y)) = f(x + c, y + d) = (x + c + a, y + d + b).$$

This is a translation in the direction of $(c + a, d + b)$. It can easily be verified that the set of all translations in $\mathbb{R}^2$ forms an abelian group, $(\mathrm{T}(2), \circ)$, under composition. The identity is the identity transformation $1_{\mathbb{R}^2}: \mathbb{R}^2 \to \mathbb{R}^2$, and the inverse of the translation in the direction (a, b) is the translation in the opposite direction $(-a, -b)$.

A function $f: X \to Y$ is called **injective** or **one-to-one** if $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. In other words, an injective function never takes two different points to the same point. The function $f: X \to Y$ is called **surjective** or

onto if for any $y \in Y$, there exists $x \in X$ with $y = f(x)$, that is, if the image $f(X)$ is the whole set Y. A **bijective** function or **one-to-one correspondence** is a function that is both injective and surjective. A **permutation** or **symmetry** of a set X is a bijection from X to itself.

Lemma 3.2. If $f: X \to Y$ and $g: Y \to Z$ are two functions, then:

(i) If f and g are injective, $g \circ f$ is injective.
(ii) If f and g are surjective, $g \circ f$ is surjective.
(iii) If f and g are bijective, $g \circ f$ is bijective.

Proof. (i) Suppose that $(g \circ f)(x_1) = (g \circ f)(x_2)$. Then $g(f(x_1)) = g(f(x_2))$ so, since g is injective, $f(x_1) = f(x_2)$. Since f is also injective, $x_1 = x_2$, proving that $g \circ f$ is injective.

(ii) Let $z \in Z$. Since g is surjective, there exists $y \in Y$ with $g(y) = z$, and since f is also surjective, there exists $x \in X$ with $f(x) = y$. Hence $(g \circ f)(x) = g(f(x)) = g(y) = z$, so $g \circ f$ is surjective.

(iii) This follows from parts (i) and (ii). □

The following theorem gives a necessary and sufficient condition for a function to have an inverse.

Theorem 3.3. Inversion Theorem. The function $f: X \to Y$ has an inverse if and only if f is bijective.

Proof. Suppose that $h: Y \to X$ is an inverse of f. The function f is injective because if $f(x_1) = f(x_2)$, it follows that $(h \circ f)(x_1) = (h \circ f)(x_2)$, and so $x_1 = x_2$. The function f is surjective because if y is any element of Y and $x = h(y)$, it follows that $f(x) = f(h(y)) = y$. Therefore, f is bijective.

Conversely, suppose that f is bijective. We define the function $h: Y \to X$ as follows. For any $y \in Y$, there exists $x \in X$ with $y = f(x)$. Since f is injective, there is only one such element x. Define $h(y) = x$. This function h is an inverse to f because $f(h(y)) = f(x) = y$, and $h(f(x)) = h(y) = x$. □

Theorem 3.4. If $S(X)$ is the set of bijections from any set X to itself, then $(S(X), \circ)$ is a group under composition. This group is called the **symmetric group** or **permutation group** of X.

Proof. It follows from Lemma 3.1 that the composition of two bijections is a bijection; thus $S(X)$ is closed under composition. The composition of functions is always associative, and the identity of $S(X)$ is the identity function $1_X: X \to X$. The inversion theorem (Theorem 3.3) proves that any bijective function $f \in S(X)$ has an inverse $f^{-1} \in S(X)$. Therefore, $(S(X), \circ)$ satisfies all the axioms for a group. □

TABLE 3.2. Symmetry Group of $\{a, b\}$

$\circ$	1_X	f
1_X	1_X	f
f	f	1_X

For example, if $X = \{a, b\}$ is a two-element set, the only bijections from X to itself are the identity 1_X and the symmetry $f: X \to X$, defined by $f(a) = b$, $f(b) = a$, that interchanges the two elements. The use of the term *symmetry* to describe the bijection f agrees with one of our everyday uses of the word. In the phrase "the boolean expression $(a \wedge b) \vee (a' \wedge b')$ is symmetrical in a and b" we mean that the expression is unchanged when we interchange a and b. The symmetric group of X, $S(X) = \{1_X, f\}$ and its group table is given in Table 3.2. The composition $f \circ f$ interchanges the two elements a and b twice; thus it is the identity.

Since the composition of functions is not generally commutative, $S(X)$ is not usually an abelian group. Consider the elements f and g in the permutation group of $\{1, 2, 3\}$, where $f(1) = 2$, $f(2) = 3$, $f(3) = 1$ and $g(1) = 1$, $g(2) = 3$, $g(3) = 2$. Then $f \circ g(1) = 2$, $f \circ g(2) = 1$, $f \circ g(3) = 3$, while $g \circ f(1) = 3$, $g \circ f(2) = 2$, $g \circ f(3) = 1$; hence $f \circ g \neq g \circ f$, and $S(\{1, 2, 3\})$ is not abelian.

A nonsingular linear transformation of the plane is a bijective function of the form $f: \mathbb{R}^2 \to \mathbb{R}^2$, where $f(x, y) = (a_{11}x + a_{12}y, a_{21}x + a_{22}y)$ with the determinant $a_{11}a_{22} - a_{12}a_{21} \neq 0$. It can be verified that the composition of two such linear transformations is again of the same type. The set of all nonsingular linear transformations, L, forms a non-abelian group $(L, \circ)$.

Besides talking about the symmetries of a distinct set of elements, we often refer, in everyday language, to a geometric object or figure as being symmetrical. We now make this notion more mathematically precise.

If F is a figure in the plane or in space, a **symmetry** of the figure F or **isometry** of F is a bijection $f: F \to F$ which preserves distances; that is, for all points $p, q \in F$, the distance from $f(p)$ to $f(q)$ must be the same as the distance from p to q.

One can visualize this operation by imagining F to be a solid object that can be picked up and turned in some manner so that it assumes a configuration identical to the one it had in its original position. For example, the design on the left of Figure 3.1 has two symmetries: the identity and a half turn about a vertical axis, called an axis of symmetry. The design in the center of Figure 3.1

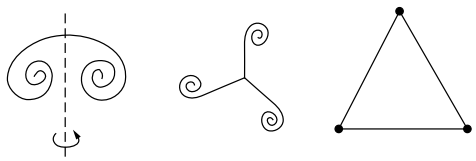


Figure 3.1. Symmetrical designs.

has three symmetries: the identity and rotations of one-third and two-thirds of a revolution about its center.

However, both the one-third rotation and interchanging two vertices are symmetries of the equilateral triangle on the right in Figure 3.1, but there is a subtle difference: The rotation can be performed as a physical motion *within* the plane of the triangle (and so is called a **proper symmetry** or a **proper rotation**), while the reflection can only be accomplished as a physical motion by moving the triangle *outside* its plane (an **improper symmetry** or an **improper rotation**).

The set of *all* symmetries of a geometric figure forms a group under composition because the composition and inverse of two distance-preserving functions is distance preserving.

Example 3.5. Write down the table for the group of symmetries of a rectangle with unequal sides.

Solution. Label the corners of the rectangle 1, 2, 3, and 4 as in Figure 3.2. Any symmetry of the rectangle will send corner points to corner points and so will permute the corners among themselves. Denote the (improper) symmetry obtained by reflecting the rectangle in the horizontal axis through the center, by a; then $a(1) = 4, a(2) = 3, a(3) = 2$, and $a(4) = 1$. This symmetry can also be considered as a rotation of the rectangle through half a revolution about this horizontal axis. There is a similar symmetry, b, about the vertical axis through the center. A third (proper) symmetry, c, is obtained by rotating the rectangle in its plane through half a revolution about its center. Finally, the identity map, e, is a symmetry. These are the only symmetries because it can be verified that any other bijection between the corners will not preserve distances.

The group of symmetries of the rectangle is $(\{e, a, b, c\}, \circ)$, and its table, as shown in Table 3.3, can be calculated as follows. The symmetries a, b, and c are all half turns, so $a \circ a$, $b \circ b$, and $c \circ c$ are full turns and are therefore equal to the identity. The function $a \circ b$ acts on the corner points by $a \circ b(1) = a(b(1)) = a(2) = 3, a \circ b(2) = 4$, $a \circ b(3) = 1$, and $a \circ b(4) = 2$. Therefore, $a \circ b = c$. The other products can be calculated similarly. □

This group of symmetries of a rectangle is sometimes called the **Klein 4-group**, after the German geometer Felix Klein (1849–1925).

We have seen that the group operation can be denoted by various symbols, the most common being multiplication, composition, and addition. It is conventional

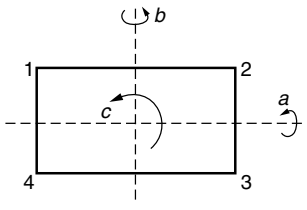


Figure 3.2. Symmetries of a rectangle.

TABLE 3.3. Symmetry Group of a Rectangle

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

to use addition only for abelian groups. Furthermore, the identity under addition is usually denoted by 0 and the inverse of a by $-a$. Hence expressions of the form $a \cdot b^{-1}$ and $a^n = a \cdots a$, in multiplicative notation, would be written as $a - b$ and $na = a + \cdots + a$, respectively, in additive notation.

In propositions and theorems concerning groups in general, it is conventional to use multiplicative notation and also to omit the dot in writing a product; therefore, $a \cdot b$ is just written as ab.

Whenever the operation in a group is clearly understood, we denote the group just by its underlying set. Therefore, the groups $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$ are usually denoted just by $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$, respectively. This should cause no confusion because $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ are not groups under multiplication (since the element 0 has no multiplicative inverse). The symmetric group of X is denoted just by $S(X)$, the operation of composition being understood. Moreover, if we refer to a group G without explicitly defining the group or the operation, it can be assumed that the operation in G is multiplication.

We now prove two propositions that will enable us to manipulate the elements of a group more easily. Recall from Proposition 2.10 that the identity of any binary operation is unique. We first show that the inverse of any element of a group is unique.

Proposition 3.6. Let $\star$ be an associative binary operation on a set S that has identity e. Then, if an element a has an inverse, this inverse is unique.

Proof. Suppose that b and c are both inverses of a; thus $a \star b = b \star a = e$, and $a \star c = c \star a = e$. Now, since e is the identity and $\star$ is associative,

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

Hence the inverse of a is unique. □

Note that if $ab = e$ in a group G with identity e, then $a^{-1} = b$ and $b^{-1} = a$. Indeed, b has an inverse b^{-1} in G, so $b^{-1} = eb^{-1} = (ab)b^{-1} = ae = a$. Similarly, $a^{-1} = b$.

Proposition 3.7. If a, b, and c are elements of a group G, then

(i) $(a^{-1})^{-1} = a$.
(ii) $(ab)^{-1} = b^{-1}a^{-1}$.
(iii) $ab = ac$ or $ba = ca$ implies that $b = c$. (cancellation law)

Proof. (i) The inverse of a^{-1} is an element b such that $a^{-1}b = ba^{-1} = e$. But a is such an element, and by Proposition 3.6 we know that the inverse is unique. Hence $(a^{-1})^{-1} = a$.

(ii) Using associativity, we have

$$(ab)(b^{-1}a^{-1}) = a((bb^{-1})a^{-1}) = a(ea^{-1}) = aa^{-1} = e.$$

Hence $b^{-1}a^{-1}$ is the unique inverse of ab.

(iii) Suppose that $ab = ac$. Then $a^{-1}(ab) = a^{-1}(ac)$, so $(a^{-1}a)b = (a^{-1}a)c$. That is, $eb = ec$ and $b = c$. Similarly, $ba = ca$ implies that $b = c$. □

Notice in part (ii) that the order of multiplication is reversed. This should be familiar from the particular case of the group of invertible $n \times n$ matrices under multiplication. A more everyday example is the operation of putting on socks and shoes. To reverse the procedure, the shoes are taken off first and then the socks.

If a is an element in any group G, we write $a^1 = a, a^2 = aa, a^3 = aaa$, and so on, just as for numbers. Hence, if $k \geqslant 1$, we define a^k to be the product of a with itself k times. Similarly, we define $a^{-k} = (a^{-1})^k$ for $k \geqslant 1$. Finally, we set a^0 to be the identity, again as for numbers. Thus we have defined a^k for every $k \in \mathbb{Z}$, and it is a routine verification that the **laws of exponents** hold:

$$a^k a^m = a^{k+m}, \quad (a^k)^m = a^{k+m} \quad \text{for all } a \in G \text{ and all } k, m \in \mathbb{Z}.$$

Moreover:

$$\text{If } ab = ba \text{ in } G, \text{ then } (ab)^k = a^k b^k \quad \text{for all } k \in \mathbb{Z}.$$

However, this need not hold if $ab \neq ba$: For example, if $a = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$ in the group of invertible 2×2 matrices, then $(ab)^2 \neq a^2b^2$.

SUBGROUPS

It often happens that some subset of a group will also form a group under the same operation. Such a group is called a *subgroup*. For example, $(\mathbb{R}, +)$ is a subgroup of $(\mathbb{C}, +)$, and the group of translations of $\mathbb{R}^2$ in Example 3.1 is a subgroup of the group of all isometries of $\mathbb{R}^2$.

If $(G, \cdot)$ is a group and H is a nonempty subset of G, then $(H, \cdot)$ is called a **subgroup** of $(G, \cdot)$ if the following conditions hold:

(i) $a \cdot b \in H$ for all $a, b \in H$. (closure)

(ii) $a^{-1} \in H$ for all $a \in H$. (existence of inverses)

Proposition 3.8. If H is a subgroup of $(G, \cdot)$, then $(H, \cdot)$ is also a group.

Proof. If H is a subgroup of $(G, \cdot)$, we show that $(H, \cdot)$ satisfies all the group axioms. The definition above implies that H is closed under the operation; that is, $\cdot$ is a binary operation on H. If $a, b, c \in H$, then $(a \cdot b) \cdot c = a \cdot (b \cdot c)$ in $(G, \cdot)$ and hence also in $(H, \cdot)$. Since H is nonempty, it contains at least one element, say h. Now $h^{-1} \in H$ and $h \cdot h^{-1}$, which is the identity, is in H. The definition of subgroup implies that $(H, \cdot)$ contains inverses. Therefore, $(H, \cdot)$ satisfies all the axioms of a group. □

Conditions (i) and (ii) are equivalent to the single condition:

(iii) $a \cdot b^{-1} \in H$ for all $a, b \in H$.

However, when H is finite, the following result shows that it is sufficient just to check condition (i).

Proposition 3.9. If H is a nonempty *finite* subset of a group G and $ab \in H$ for all $a, b \in H$, then H is a subgroup of G.

Proof. We have to show that for each element $a \in H$, its inverse is also in H. All the elements, $a, a^2 = aa, a^3 = aaa, \ldots$ belong to H so, since H is finite, these cannot all be distinct. Therefore, $a^i = a^j$ for some $1 \leqslant i < j$. By Proposition 3.7(iii), we can cancel a^i from each side to obtain $e = a^{j-1}$, where $j - i > 0$. Therefore, $e \in H$ and this equation can be written as $e = a(a^{j-i-1}) = (a^{j-i-1})a$. Hence $a^{-1} = a^{j-i-1}$, which belongs to H, since $j - i - 1 \geqslant 0$. □

In the group $(\{1, -1, i, -i\}, \cdot)$, the subset $\{1, -1\}$ forms a subgroup because this subset is closed under multiplication. In the group of translations of the plane (Example 3.1), the set of translations in the horizontal direction forms a subgroup because compositions and inverses of horizontal translations are still horizontal translations.

The group $\mathbb{Z}$ is a subgroup of $\mathbb{Q}$, $\mathbb{Q}$ is a subgroup of $\mathbb{R}$, and $\mathbb{R}$ is a subgroup of $\mathbb{C}$. (Remember that addition is the operation in all these groups.)

However, the set $\mathbb{N} = \{0, 1, 2, \ldots\}$ of nonnegative integers is a subset of $\mathbb{Z}$ but *not* a subgroup, because the inverse of 1, namely, -1, is not in $\mathbb{N}$. This example shows that Proposition 3.9 is false if we drop the condition that H be finite.

The relation of being a subgroup is transitive. In fact, for any group G, the inclusion relation between the subgroups of G is a partial order relation.

Example 3.10. Draw the poset diagram of the subgroups of the group of symmetries of a rectangle.

Solution. By looking at the table of this group in Table 3.3, we see that $\circ$ is a binary operation on $\{e, a\}$; thus $\{e, a\}$ is a subgroup. Also, $\{e, b\}$ and $\{e, c\}$ are subgroups. If a subgroup contains a and b, it must contain $a \circ b = c$, so it is the

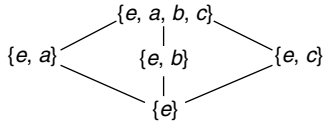


Figure 3.3. Subgroups of the group of symmetries of a rectangle.

whole group. Similarly, subgroups containing a and c or b and c must be the whole group. The poset diagram of subgroups is given in Figure 3.3. □

CYCLIC GROUPS AND DIHEDRAL GROUPS

The number of elements in a group G is written $|G|$ and is called the **order of the group**. G is called a **finite group** if $|G|$ is finite, and G is called an **infinite group** otherwise.

An important class of groups consists of those for which every element can be written as a power (positive or negative) of some fixed element. More precisely, a group $(G, \cdot)$ is called **cyclic** if there exists an element $g \in G$ such that $G = \{g^n | n \in \mathbb{Z}\}$. The element g is called a **generator** of the cyclic group.

Every cyclic group is abelian because $g^r \cdot g^s = g^{r+s} = g^s \cdot g^r$.

The group $(\{1, -1, i, -i\}, \cdot)$ is a cyclic group of order 4 generated by i because $i^0 = 1, i^1 = i, i^2 = -1, i^3 = -i, i^4 = 1, i^5 = i$, and so on. Hence the group can be written as $(\{1, i, i^2, i^3\}, \cdot)$.

In additive notation, the group $(G, +)$ is cyclic if $G = \{ng | n \in \mathbb{Z}\}$ for some $g \in G$. The group $(\mathbb{Z}, +)$ is an infinite cyclic group with generator 1 (or -1).

The **order of an element** g in a group $(G, \cdot)$ is the least positive integer r such that $g^r = e$. If no such r exists, the order of the element is said to be *infinite*.

Note the difference between the order of an *element* and the order of a *group*. We are going to find connections between these two orders and later prove Lagrange's theorem, which implies that in a finite group, the order of every element divides the order of the group.

For example, in $(\{1, -1, i, -i\}, \cdot)$, the identity 1 has order 1, -1 has order 2 because $(-1)^2 = 1$, whereas i and $-i$ both have order 4. The group has order 4.

Let $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ be the set of nonzero rational numbers. Then $(\mathbb{Q}^*, \cdot)$ is a group under multiplication. The order of the identity element 1 is 1, and the order of -1 is 2. The order of every other element is infinite, because the only solutions to $q^r = 1$ with $q \in \mathbb{Q}^*, r \geqslant 1$ are $q = \pm 1$. The group has infinite order. However, it is not cyclic, because there is no rational number r such that every nonzero rational can be written as r^n for some $n \in \mathbb{Z}$.

The next two results show how the division algorithm for integers (see Appendix 2) is used in group theory.

Proposition 3.11. Let a be an element of order r in a group G. Then for $k \in \mathbb{Z}$, $g^k = e$ if and only if r divides k.

Proof. If $k = rm, m \in \mathbb{Z}$, then $a^k = (a^r)^m = e^m = e$. Conversely, if $a^k = e$, write $k = qr + s$, where q and s are in $\mathbb{Z}$ and $0 \leqslant s < r$. Then $a^s = a^{k-qr} = a^k(a^r)^{-q} = e \cdot e^{-q} = e$. Since $0 \leqslant s < r$ and r is the *smallest* positive integer such that $a^r = e$, it follows that $s = 0$. But then $k = qr$, as required. □

Proposition 3.12. Every subgroup of a cyclic group is cyclic.

Proof. Suppose that G is cyclic with generator g and that $H \subseteq G$ is a subgroup. If $H = \{e\}$, it is cyclic with generator e. Otherwise, let $g^k \in H$ with $k \neq 0$. Since $g^{-k} = (g^k)^{-1}$ is in H, we have $g^m \in H$ for some $m > 0$, and we choose m to be the smallest such positive integer. Write $h = g^m$; we claim that h generates H. Certainly, $h^k \in H$ for every $k \in \mathbb{Z}$ because $h \in H$; we must show that *every* element a in H is a power of h. Since $a \in G$ we have $a = g^s, s \in \mathbb{Z}$. By the division algorithm, write $s = qm + r$, where $0 \leqslant r < m$. Then $a^r = g^{s-qm} = g^s(g^m)^{-q} = a \cdot h^{-q} \in H$, so $r = 0$ by the choice of m. Hence $a = g^s = g^{qm} = (g^m)^q = h^q$, as we wanted. □

For any element g in a group $(G, \cdot)$ we can look at all the powers of this element, namely, $\{g^r | r \in \mathbb{Z}\}$. This may not be the whole group, but it will be a subgroup.

Proposition 3.13. If g is any element of order k in a group $(G, \cdot)$, then $H = \{g^r | r \in \mathbb{Z}\}$ is a subgroup of order k in $(G, \cdot)$. This is called the **cyclic subgroup generated by** g.

Proof. We first check that H is a subgroup of $(G, \cdot)$. This follows from the fact that $g^r \cdot g^s = g^{r+s} \in H$ and $(g^r)^{-1} = g^{-r} \in H$ for all $r, s \in \mathbb{Z}$.

If the order of the element g is infinite, we show that the elements g^r are all distinct. Suppose that $g^r = g^s$, where $r > s$. Then $g^{r-s} = e$ with $r - s > 0$, which contradicts the fact that g has infinite order. In this case, $|H|$ is infinite.

If the order of the element g is k, which is finite, we show that $H = \{g^0 = e, g^1, g^2, \ldots, g^{k-1}\}$. Suppose that $g^r = g^s$, where $0 \leqslant s < r \leqslant k - 1$. Multiply both sides by g^{-s} so that $g^{r-s} = e$ with $0 < r - s < k$. This contradicts the fact that k is the order of g. Hence the elements $g^0, g^1, g^2, \ldots, g^{k-1}$ are all distinct. For any other element, g^t, we can write $t = qk + r$, where $0 \leqslant r < k$ by the division algorithm. Hence

$$g^t = g^{qk+r} = (g^k)^q(g^r) = (e^q)(g^r) = g^r.$$

Hence $H = \{g^0, g^1, g^2, \ldots, g^{k-1}\}$ and $|H| = k$. □

For example, in $(\mathbb{Z}, +)$, the subgroup generated by 3 is $\{\ldots, -3, 0, 3, 6, 9, \ldots\}$, an infinite subgroup that we write as $3\mathbb{Z} = \{3r | r \in \mathbb{Z}\}$.

Theorem 3.14. If the finite group G is of order n and has an element g of order n, then G is a cyclic group generated by g.

Proof. From the previous proposition we know that H, the subgroup of G generated by g, has order n. Therefore, H is a subset of the finite set G with the same number of elements. Hence $G = H$ and G is a cyclic group generated by g. □

Example 3.15. Show that the Klein 4-group of symmetries of a rectangle, described in Example 3.5, is not cyclic.

Solution. In the Klein 4-group, the identity has order 1, whereas all the other elements have order 2. As it has no element of order 4, it cannot be cyclic. □

All the elements of the Klein 4-group can be written in terms of a and b. We therefore say that this group can be **generated** by the two elements a and b.

Example 3.16. Show that the group of proper rotations of a regular n-gon in the plane is a cyclic group of order n generated by a rotation of $2\pi/n$ radians. This group is denoted by C_n.

Solution. This is the group of those symmetries of the regular n-gon that can be performed in the plane, that is, without turning the n-gon over.

Label the vertices 1 through n as in Figure 3.4. Under any symmetry, the center must be fixed, and the vertex 1 can be taken to any of the n vertices. The image of 1 determines the rotation; hence the group is of order n.

Let g be the counterclockwise rotation of the n-gon through $2\pi/n$. Then g has order n, and by Theorem 3.14, the group is cyclic of order n. Hence $C_n = \{e, g, g^2, \ldots, g^{n-1}\}$. □

Let us now consider the group of *all* symmetries (both proper and improper rotations) of the regular n-gon. We call this group the **dihedral group** and denote it by D_n.

Example 3.17. Show that the dihedral group, D_n, is of order $2n$ and is not cyclic.

Solution. Label the vertices 1 to n in a counterclockwise direction around the n-gon. Let g be a counterclockwise rotation through $2\pi/n$, and let h be the improper rotation of the n-gon about an axis through the center and vertex 1, as indicated in Figure 3.5. The element g generates the group C_n, which is a cyclic subgroup of D_n. The element h has order 2 and generates a subgroup $\{e, h\}$.

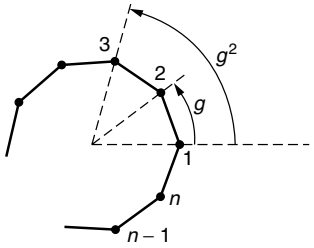


Figure 3.4. Elements of C_n.

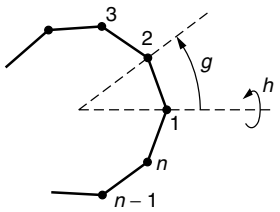


Figure 3.5. Elements of D_n.

Any symmetry will fix the origin and is determined by the image of two adjacent vertices, say 1 and 2. The vertex 1 can be taken to any of the n vertices, and then 2 must be taken to one of the two vertices adjacent to the image of 1. Hence D_n has order $2n$.

If the image of 1 is $r+1$, the image of 2 must be $r+2$ or r. If the image of 2 is $r+2$, the symmetry is g^r. If the image of 2 is r, the symmetry is $g^r h$. Figure 3.6 shows that the symmetries $g^r h$ and hg^{-r} have the same effect and therefore imply the relation $g^r h = hg^{-r} = hg^{n-r}$.

Hence the dihedral group is

$$D_n = \{e, g, g^2, \ldots, g^{n-1}, h, gh, g^2h, \ldots, g^{n-1}h\}.$$

Note that if $n \geqslant 3$, then $gh \neq hg$; thus D_n is a *noncommutative* group. Therefore, this group cannot be cyclic. □

D_2 can be defined as the symmetries of the figure in Figure 3.7. Hence $D_2 = \{e, g, h, gh\}$, and each nonidentity element has order 2.

Example 3.18. Draw the group table for C_4 and D_4.

Solution. D_4 is the group of symmetries of the square, and its table, which is calculated using the relation $g^r h = hg^{4-r}$, is given in Table 3.4. For example, $(g^2h)(gh) = g^2(hg)h = g^2(g^3h)h = g^5h^2 = g$. Since C_4 is a subgroup of D_4, the table for C_4 appears inside the dashed lines in the top left corner. □

Note that the order of each of the elements h, gh, g^2h, and g^3h in D_4 is 2. In general, the element $g^r h$ in D_n is a reflection in the line through the center

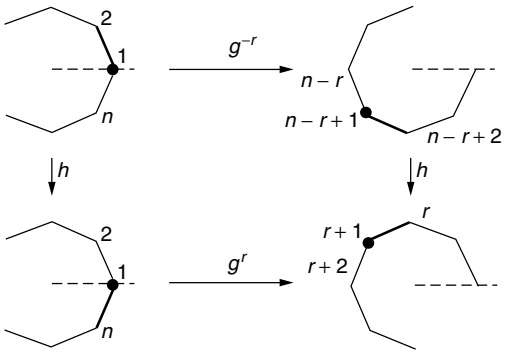


Figure 3.6. Relation $g^r h = hg^{-r}$ in D_n.



Figure 3.7. Symmetries of a 2-gon.

TABLE 3.4. Group D_4

$\cdot$	e	g	g^2	g^3	h	gh	g^2h	g^3h
e	e	g	g^2	g^3	h	gh	g^2h	g^3h
g	g	g^2	g^3	e	gh	g^2h	g^3h	h
g^2	g^2	g^3	e	g	g^2h	g^3h	h	gh
g^3	g^3	e	g	g^2	g^3h	h	gh	g^2h
h	h	g^3h	g^2h	gh	e	g^3	g^2	g
gh	gh	h	g^3h	g^2h	g	e	g^3	g^2
g^2h	g^2h	gh	h	g^3h	g^2	g	e	g^3
g^3h	g^3h	g^2h	gh	h	g^3	g^2	g	e

of the n-gon bisecting the angle between vertices 1 and $r+1$. Therefore, $g^r h$ always has order 2.

MORPHISMS

Recall that a morphism between two algebraic structures is a function that preserves their operations. For instance, in Example 3.5, each element of the group K of symmetries of the rectangle induces a permutation of the vertices 1, 2, 3, 4. This defines a function $f: K \to S(\{1, 2, 3, 4\})$ with the property that the composition of two symmetries of the rectangle corresponds to the composition of permutations of the set $\{1, 2, 3, 4\}$. Since this function preserves the operations, it is a morphism of groups.

Two groups are *isomorphic* if their structures are essentially the same. For example, the group tables of the cyclic group C_4 and $(\{1, -1, i, -i\}, \cdot)$ would be identical if we replaced a rotation through $n\pi/2$ by i^n. We would therefore say that $(C_4, \circ)$ and $(\{1, -1, i, -1\}, \cdot)$ are isomorphic.

If $(G, \cdot)$ and $(H, \cdot)$ are two groups, the function $f: G \to H$ is called a **group morphism** if

$$f(a \cdot b) = f(a) \cdot f(b) \quad \text{for all } a, b \in G.$$

If the groups have different operations, say they are $(G, \cdot)$ and $(H, \star)$, the condition would be written as

$$f(a \cdot b) = f(a) \star f(b).$$

We often use the notation $f: (G, \cdot) \to (H, \star)$ for such a morphism. Many authors use *homomorphism* instead of *morphism* but we prefer the simpler terminology.

A **group isomorphism** is a bijective group morphism. If there is an isomorphism between the groups $(G, \cdot)$ and $(H, \star)$, we say that $(G, \cdot)$ and $(H, \star)$ are **isomorphic** and write $(G, \cdot) \cong (H, \star)$.

If G and H are any two groups, the trivial function that maps every element of G to the identity of H is always a morphism. If $i: \mathbb{Z} \to \mathbb{Q}$ is the inclusion map, i is a group morphism from $(\mathbb{Z}, +)$ to $(\mathbb{Q}, +)$. In fact, if H is a subgroup of G, the inclusion map $H \to G$ is always a group morphism.

Let $f: \mathbb{Z} \to \{1, -1\}$ be the function defined by $f(n) = 1$ if n is even, and $f(n) = -1$ if n is odd. Then it can be verified that $f(m + n) = f(m) \cdot f(n)$ for any $m, n \in \mathbb{Z}$, so this defines a group morphism $f: (\mathbb{Z}, +) \to (\{1, -1\}, \cdot)$.

Let $GL(2, \mathbb{R})$ be the set of 2×2 invertible real matrices. The one-to-one correspondence between the set, L, of invertible linear transformations of the plane and the 2×2 coefficient matrices is an isomorphism between the groups $(L, \circ)$ and $(GL(2, \mathbb{R}), \cdot)$.

Isomorphic groups share exactly the same properties, and we sometimes identify the groups via the isomorphism and give them the same name. If $f: G \to H$ is an isomorphism between finite groups, the group table of H is the same as that of G, when each element $g \in G$ is replaced by $f(g) \in H$.

Besides preserving the operations of a group, the following result shows that morphisms also preserve the identity and inverses.

Proposition 3.19. Let $f: G \to H$ be a group morphism, and let e_G and e_H be the identities of G and H, respectively. Then

(i) $f(e_G) = e_H$.
(ii) $f(a^{-1}) = f(a)^{-1}$ for all $a \in G$.

Proof. (i) Since f is a morphism, $f(e_G)f(e_G) = f(e_G \cdot e_G) = f(e_G) = f(e_G)e_H$. Hence (i) follows by cancellation in H (Proposition 3.7).

(ii) $f(a) \cdot f(a^{-1}) = f(a \cdot a^{-1}) = f(e_G) = e_H$ by (i). Hence $f(a^{-1})$ is the unique inverse of $f(a)$; that is $f(a^{-1}) = f(a)^{-1}$. □

Theorem 3.20. Cyclic groups of the same order are isomorphic.

Proof. Let $G = \{g^r | r \in \mathbb{Z}\}$ and $H = \{h^r | r \in \mathbb{Z}\}$ be cyclic groups. If G and H are infinite, then g has infinite order, so for $r, s \in \mathbb{Z}$, $g^r = g^s$ if and only if $r = s$ (see Proposition 3.13). Hence the function $f: G \to H$ defined by $f(g^r) = h^r, r \in \mathbb{Z}$, is a bijection, and

$$f(g^r g^s) = f(g^{r+s}) = h^{r+s} = h^r h^s = f(g^r) f(g^s)$$

for all $r, s \in \mathbb{Z}$, so f is a group isomorphism.

If $|G| = n = |H|$, then $G = \{e, g, g^2, \ldots, g^{n-1}\}$, where these powers of g are all distinct (see the proof of Proposition 3.13). Similarly, $H = \{e, h, h^2, \ldots, h^{n-1}\}$. Then the function $f: G \to H$ defined by $f(g^r) = h^r$, is again a bijection. To see that it is a morphism, suppose that $0 \leqslant r, s \leqslant n - 1$, and let $r + s = kn + l$, where $0 \leqslant l \leqslant n - 1$. Then

$$f(g^r \cdot g^s) = f(g^{r+s}) = f(g^{kn+l}) = f((g^n)^k \cdot g^l) = f(e^k \cdot g^l) = f(g^l) = h^l$$

and

$$f(g^r) \cdot f(g^s) = h^r \cdot h^s = h^{r+s} = h^{kn+l} = (h^n)^k \cdot h^l = e^k \cdot h^l = h^l,$$

so f is an isomorphism. □

Hence every cyclic group is isomorphic to either $(\mathbb{Z}, +)$ or $(C_n, \cdot)$ for some n. In the next chapter, we see that another important class of cyclic groups consists of the integers modulo n, $(\mathbb{Z}_n, +)$. Of course, the theorem above implies that $(\mathbb{Z}_n, +) \cong (C_n, \cdot)$.

Any morphism, $f: G \to H$, from a *cyclic* group G to any group H is determined just by the image of a generator. If g generates G and $f(g) = h$, it follows from the definition of a morphism that $f(g^r) = f(g)^r = h^r$ for all $r \in \mathbb{Z}$.

Proposition 3.21. Corresponding elements under a group isomorphism have the same order.

Proof. Let $f: G \to H$ be an isomorphism, and let $f(g) = h$. Suppose that g and h have orders m and n, respectively, where m is finite. Then $h^m = f(g)^m = f(g^m) = f(e) = e$. So n is also finite, and $n \leqslant m$, since n is the *least* positive integer with the property $h^n = e$.

On the other hand, if n is finite then $f(g^n) = f(g)^n = h^n = e = f(e)$. Since f is bijective, $g^n = e$, and hence m is finite and $m \leqslant n$.

Therefore, either m and n are both finite and $m = n$, or m and n are both infinite. □

Example 3.22. Is D_2 isomorphic to C_4 or the Klein 4-group of symmetries of a rectangle?

Solution. Compare the orders of the elements given in Table 3.5. By Proposition 3.21 we see that D_2 cannot be isomorphic to C_4 but could possibly be isomorphic to the Klein 4-group.

In the Klein 4-group we can write $c = a \circ b$ and we can obtain a bijection, f, from D_2 to the Klein 4-group, by defining $f(g) = a$ and $f(h) = b$. Table 3.6 for the two groups show that this is an isomorphism. □

TABLE 3.5

D_2		C_4		Klein 4-Group	
Element	Order	Element	Order	Element	Order
e	1	e	1	e	1
g	2	g	4	a	2
h	2	g^2	2	b	2
gh	2	g^3	4	c	2

TABLE 3.6. Isomorphic Groups

Group D_2					Klein 4-Group				
$\cdot$	e	g	h	gh	$\circ$	e	a	b	c
e	e	g	h	gh	e	e	a	b	c
g	g	e	gh	h	a	a	e	c	b
h	h	gh	e	g	b	b	c	e	a
gh	gh	h	g	e	c	c	b	a	e

PERMUTATION GROUPS

A *permutation* of n elements is a bijective function from the set of the n elements to itself. The permutation groups of two sets, with the same number of elements, are isomorphic. We denote the permutation group of $X = \{1, 2, 3, \ldots, n\}$ by $(S_n, \circ)$ and call it the **symmetric group on n elements**. Hence $S_n \cong S(Y)$ for any n element set Y.

Proposition 3.23. $|S_n| = n!$

Proof. The order of S_n is the number of bijections from $\{1, 2, \ldots, n\}$ to itself. There are n possible choices for the image of 1 under a bijection. Once the image of 1 has been chosen, there are $n - 1$ choices for the image of 2. Then there are $n - 2$ choices for the image of 3. Continuing in this way, we see that $|S_n| = n(n-1)(n-2)\cdots 2 \cdot 1 = n!$ □

If $\pi: \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ is a permutation, we denote it by

$$\begin{pmatrix} 1 & 2 & \cdots & \mathrm{n} \\ \pi(1) & \pi(2) & \cdots & \pi(n) \end{pmatrix}.$$

For example, the permutation of $\{1, 2, 3\}$ that interchanges 1 and 3 is written $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$. We think of this as

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

We can write $S_2 = \left\{ \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \right\}$, which has two elements and

$$S_3 = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \right\},$$

which is of order $3! = 6$.

If $\pi, \rho \in S_n$ are two permutations, their product $\pi \circ \rho$ is the permutation obtained by applying ρ first and then π. This agrees with our notion of composition of functions because $(\pi \circ \rho)(x) = \pi(\rho(x))$. (However, the reader should be aware that some authors use the opposite convention in which π is applied first and then ρ.)

Example 3.24. If $\pi = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$ and $\rho = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$ are two elements of S_3, calculate $\pi \circ \rho$ and $\rho \circ \pi$.

Solution. $\pi \circ \rho = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$. To calculate this, we start at the right and trace the image of each element under the composition.

$$\begin{bmatrix} 1 & 2 & 3 \\ & & \downarrow \\ 3 & 1 & 2 \end{bmatrix} \circ \begin{bmatrix} 1 & 2 & 3 \\ \downarrow & & \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ \downarrow & & \\ 2 & & \end{bmatrix}$$

Under ρ, 1 is mapped to 3, and under π, 3 is mapped to 2; thus under $\pi \circ \rho$, 1 is mapped to 2. Tracing the images of 2 and 3, we see that

$$\pi \circ \rho = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

In a similar way we can show that

$$\rho \circ \pi = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

Note that $\pi \circ \rho \neq \rho \circ \pi$ and so S_3 is not commutative. □

The permutation $\pi = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$ has the effect of moving the elements around in a cycle. This is called a *cycle of length 3*, and we write this as $\begin{pmatrix} 1 & 3 & 2 \end{pmatrix}$. We think of this as $(1 \to 3 \to 2)$. The permutation π could also be written as $\begin{pmatrix} 3 & 2 & 1 \end{pmatrix}$ or $\begin{pmatrix} 2 & 1 & 3 \end{pmatrix}$ in cycle notation.

In general, if $a_1, a_2, \ldots, a_r$ are distinct elements of $\{1, 2, 3, \ldots, n\}$, the permutation $\pi \in S_n$, defined by

$$\pi(a_1) = a_2, \quad \pi(a_2) = a_3, \ldots, \pi(a_{r-1}) = a_r, \quad \pi(a_r) = a_1$$

and $\pi(x) = x$ if $x \notin \{a_1, a_2, \ldots, a_r\}$, is called a **cycle of length** r or an r-**cycle**. We denote it by $(a_1 a_2 \cdots a_r)$.

Note that the value of n does not appear in the cycle notation. For example, $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 4 & 2 \end{pmatrix}$, is a 4-cycle in S_4, whereas $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 4 & 2 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 3 & 4 & 2 \end{pmatrix}$ is a 4-cycle in S_6, and $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 5 & 3 & 2 & 4 & 6 \end{pmatrix} = \begin{pmatrix} 2 & 5 & 4 \end{pmatrix}$, is a 3-cycle in S_6.

Proposition 3.25. An r-cycle in S_n has order r.

Proof. If $\pi = (a_1 a_2 \cdots a_r)$ is an r-cycle in S_n, then $\pi(a_1) = a_2, \pi^2(a_1) = a_3, \pi^3(a_1) = a_4, \ldots$ and $\pi^r(a_1) = a_1$. Similarly, $\pi^r(a_i) = a_i$ for $i = 1, 2, \ldots, r$.

Since π^r fixes all the other elements, it is the identity permutation. But none of the permutations $\pi, \pi^2, \ldots, \pi^{r-1}$ equal the identity permutation because they all move the element a_1. Hence the order of π is r. □

Example 3.26. Write down $\pi = \begin{pmatrix} 1 & 3 & 4 & 2 \end{pmatrix}$, $\rho = \begin{pmatrix} 1 & 3 \end{pmatrix}$, and $\sigma = \begin{pmatrix} 1 & 2 \end{pmatrix} \circ \begin{pmatrix} 3 & 4 \end{pmatrix}$ as permutations in S_4. Calculate $\pi \circ \rho \circ \sigma$.

Solution.

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

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

We can either calculate a product of cycles from the permutation representation or we can use the cycle representation directly. Let us calculate $\pi \circ \rho \circ \sigma$ from their cycles. Remember that a cycle in S_4 is a bijection from $\{1, 2, 3, 4\}$ to itself, and a product of cycles is a composition of functions. In calculating such a composition, we begin at the right and work our way left. Consider the effect of $\pi \circ \rho \circ \sigma$ on each of the elements 1, 2, 3, and 4.

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

$$1 \longleftarrow 2 \longleftarrow 2 \longleftarrow 1 \longleftarrow 1$$

$$4 \longleftarrow 3 \longleftarrow 1 \longleftarrow 2 \longleftarrow 2$$

$$2 \longleftarrow 4 \longleftarrow 4 \longleftarrow 4 \longleftarrow 3$$

$$3 \longleftarrow 1 \longleftarrow 3 \longleftarrow 3 \longleftarrow 4$$

For example, 2 is left unchanged by $\begin{pmatrix} 3 & 4 \end{pmatrix}$; then 2 is sent to 1 under $\begin{pmatrix} 1 & 2 \end{pmatrix}$, 1 is sent to 3 under $\begin{pmatrix} 1 & 3 \end{pmatrix}$, and finally, 3 is sent to 4 under $\begin{pmatrix} 1 & 3 & 4 & 2 \end{pmatrix}$. Hence $\pi \circ \rho \circ \sigma$ sends 2 to 4. The permutation $\pi \circ \rho \circ \sigma$ also sends 4 to 3, 3 to 2, and fixes 1. Therefore, $\pi \circ \rho \circ \sigma = \begin{pmatrix} 2 & 4 & 3 \end{pmatrix}$. □

Permutations that are not cycles can be split up into two or more cycles as follows. If π is a permutation in S_n and $a \in \{1, 2, 3, \ldots, n\}$, the **orbit** of a under π consists of the distinct elements $a, \pi(a), \pi^2(a), \pi^3(a), \ldots$. We can split a permutation up into its different orbits, and each orbit will give rise to a cycle.

Consider the permutation $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 2 & 8 & 1 & 5 & 7 & 6 & 4 \end{pmatrix} \in S_8$. Here $\pi(1) = 3, \pi^2(1) = \pi(3) = 8, \pi^3(1) = 4$, and $\pi^4(1) = 1$; thus the orbit of 1 is $\{1, 3, 8, 4\}$. This is also the orbit of 3, 4, and 8. This orbit gives rise to the cycle $\begin{pmatrix} 1 & 3 & 8 & 4 \end{pmatrix}$. Since π leaves 2 and 5 fixed, their orbits are $\{2\}$ and $\{5\}$. The orbit of 6 and 7 is $\{6, 7\}$, which gives rise to the 2-cycle $\begin{pmatrix} 6 & 7 \end{pmatrix}$. We can picture the orbits and their corresponding cycles as in Figure 3.8.

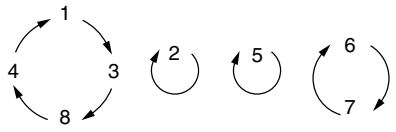


Figure 3.8. Disjoint cycle decomposition.

It can be verified that $\pi = \begin{pmatrix} 1 & 3 & 8 & 4 \end{pmatrix} \circ (2) \circ (5) \circ \begin{pmatrix} 6 & 7 \end{pmatrix}$. Since no number is in two different cycles, these cycles are called **disjoint**. If a permutation is written as a product of disjoint cycles, it does not matter in which order we write the cycles. We could write $\pi = (5) \circ \begin{pmatrix} 6 & 7 \end{pmatrix} \circ (2) \circ \begin{pmatrix} 1 & 3 & 8 & 4 \end{pmatrix}$. When writing down a product of cycles, we often omit the 1-cycles and write $\pi = \begin{pmatrix} 1 & 3 & 8 & 4 \end{pmatrix} \circ \begin{pmatrix} 6 & 7 \end{pmatrix}$. The identity permutation is usually just written as (1).

Proposition 3.27. Every permutation can be written as a product of disjoint cycles.

Proof. Let π be a permutation and let $\gamma_1, \ldots, \gamma_k$ be the cycles obtained as described above from the orbits of π. Let a_1 be any number in the domain of π, and let $\pi(a_1) = a_2$. If γ_i is the cycle containing a_1, we can write $\gamma_i = (a_1 a_2 \cdots a_r)$; the other cycles will not contain any of the elements $a_1, a_2, \ldots, a_r$ and hence will leave them all fixed. Therefore, the product $\gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_k$ will map a_1 to a_2, because the only cycle to move a_1 or a_2 is γ_i. Hence $\pi = \gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_k$, because they both have the same effect on all the numbers in the domain of π. □

Corollary 3.28. The order of a permutation is the least common multiple of the lengths of its disjoint cycles.

Proof. If π is written in terms of disjoint cycles as $\gamma_1 \circ \gamma_2 \circ \cdots \circ \gamma_k$, the order of the cycles can be changed because they are disjoint. Therefore, for any integer m, $\gamma^m = \gamma_1^m \circ \gamma_2^m \circ \cdots \circ \gamma_k^m$. Because the cycles are disjoint, this is the identity if and only if γ_i^m is the identity for each i. The least such integer is the least common multiple of the orders of the cycles. □

Example 3.29. Find the order of the permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 5 & 8 & 7 & 1 & 4 & 6 & 2 \end{pmatrix}.$$

Solution. We can write this permutation in terms of disjoint cycles as

$$\pi = \begin{pmatrix} 1 & 3 & 8 & 2 & 5 \end{pmatrix} \circ \begin{pmatrix} 4 & 7 & 6 \end{pmatrix}.$$

Hence the order of π is $\text{lcm}\,(5, 3) = 15$. Of course, we could calculate $\pi^2, \pi^3, \pi^4, \ldots$ until we obtained the identity, but this would take much longer. □

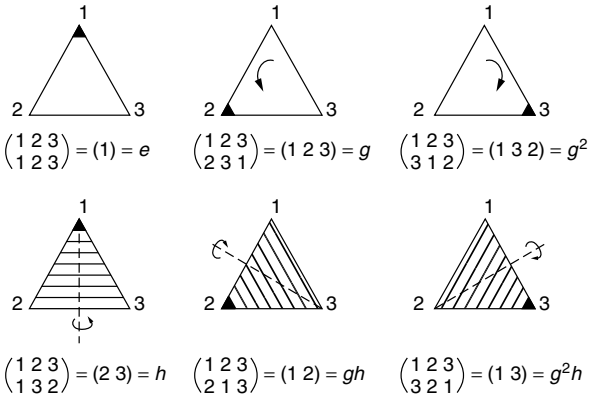


Figure 3.9. Symmetries of an equilateral triangle.

TABLE 3.7. Group S_3

$\circ$	(1)	(123)	(132)	(23)	(12)	(13)
(1)	(1)	(123)	(132)	(23)	(12)	(13)
(123)	(123)	(132)	(1)	(12)	(13)	(23)
(132)	(132)	(1)	(123)	(13)	(23)	(12)
(23)	(23)	(13)	(12)	(1)	(132)	(123)
(12)	(12)	(23)	(13)	(123)	(1)	(132)
(13)	(13)	(12)	(23)	(132)	(123)	(1)

Example 3.30. Show that D_3 is isomorphic to S_3 and write out the table for the latter group.

Solution. D_3 is the group of symmetries of an equilateral triangle, and any symmetry induces a permutation of the vertices. This defines a function $f: D_3 \to S_3$. If $\sigma, \tau \in D_3$, then $f(\sigma \circ \tau)$ is the induced permutation on the vertices, which is the same as $f(\sigma) \circ f(\tau)$. Hence f is a morphism. Figure 3.9 illustrates the six elements of D_3 and their corresponding permutations. We shade the underside of the triangle and mark the corner near vertex 1 to illustrate how the triangle moves. To visualize this, imagine a triangular jigsaw puzzle piece and consider all possible ways of fitting this piece into a triangular hole. Any proper rotation will leave the white side uppermost, whereas an improper rotation will leave the shaded side uppermost.

The six permutations are all distinct; thus f is a bijection and an isomorphism between D_3 and S_3. The group table for S_3 is given in Table 3.7. □

EVEN AND ODD PERMUTATIONS

We are going to show that every permutation can be given a parity, even or odd. The definition derives from an action of each permutation σ in S_n on a

polynomial $f(x_1, x_2, \ldots, x_n)$ in n variables by permuting the variables:

$$\sigma f(x_1, x_2, \ldots, x_n) = f(x_{\sigma 1}, x_{\sigma 2}, \ldots, x_{\sigma n}).$$

For example, if $\sigma = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$ in S_4 and $f(x_1, x_2, x_3, x_4) = 2x_1x_4 - 3x_2^2 + x_2x_3^3$, then $\sigma f = 2x_2x_4 - 3x_3^2 + x_3x_1^3$.

Our use of this action involves a particular polynomial $D = D(x_1, x_2, \ldots, x_n)$ called the **discriminant**, defined to be the product of all terms $(x_i - x_j)$, where $i < j$. More formally,

$$D = \prod_{0 \leqslant i < j \leqslant n} (x_i - x_j).$$

For example, if $n = 3$, then $D = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$. Given a permutation $\sigma \in S_n$, we have

$$\sigma D = \prod_{0 \leqslant i < j \leqslant n} (x_{\sigma i} - x_{\sigma j}).$$

Thus if $n = 3$ and $\sigma = (12) \in S_3$, then $\sigma D = (x_2 - x_1)(x_2 - x_3)(x_1 - x_3) = -D$. In fact, $\sigma D = \pm D$ for every $\sigma \in S_n$, and we say that

$$\sigma \text{ is } \textbf{even} \text{ if } \sigma D = D \quad \text{and} \quad \sigma \text{ is } \textbf{odd} \text{ if } \sigma D = -D.$$

We are going to determine an easy method for deciding which is the case. A 2-cycle is called a **transposition**, and surprisingly, much of the discussion centers around determining the parity of these transpositions. We are going to show that every transposition is odd.

Let D denote the discriminant in n variables $x_1, x_2, \ldots, x_n$, and define

$$D_{k/m} = \text{the product of all terms in } D \text{ involving } x_k, \text{ except } (x_k - x_m)$$
$$D_{k,m} = \text{the product of all terms in } D \text{ involving neither } x_k \text{ nor } x_m.$$

For example, if $n = 5$, we have

$$D_{2/4} = (x_1 - x_2)(x_2 - x_3)(x_2 - x_5)$$
$$D_{4/2} = (x_1 - x_4)(x_3 - x_4)(x_4 - x_5)$$
$$D_{2,4} = (x_1 - x_3)(x_1 - x_5)(x_3 - x_5).$$

Then D factors as follows:

$$D = (x_k - x_m) D_{k/m} D_{m/k} D_{k,m}.$$

Now fix a transposition $\tau = (k\ m)$ in S_n, where $k < m$. Since τ interchanges k and m, we see that

$$\tau D_{k/m} = u D_{m/k} \quad \text{where } u = 1 \text{ or } u = -1.$$

Since τ^2 is the identity permutation, we have

$$D_{k/m} = \tau^2 D_{k/m} = \tau(\tau D_{k/m}) = \tau(u D_{m/k}) = u(\tau D_{m/k}).$$

Because $u^2 = 1$, it follows that

$$\tau D_{m/k} = u D_{k/m}.$$

Since $\tau D_{k,m} = D_{k,m}$, applying τ to D gives

$$\begin{aligned} \tau D &= \tau(x_k - x_m) \cdot \tau D_{k/m} \cdot \tau D_{m/k} \cdot \tau D_{k,m} \\ &= (x_m - x_k) \cdot u D_{m/k} \cdot u D_{k/m} \cdot D_{k,m} \\ &= -D \end{aligned}$$

because $u^2 = 1$. Hence τ is odd, and we have proved:

Proposition 3.31. Every transposition is odd.

With this we can determine the parity of an arbitrary permutation σ in S_n. The idea is to factor σ as a product of transpositions, and we begin with the cycles. The proof of the next result is a straightforward verification.

Proposition 3.32. Every r-cycle is a product of $r - 1$ transpositions (not necessarily disjoint); in fact,

$$\begin{pmatrix} a_1 & a_2 & \cdots & a_r \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \circ \begin{pmatrix} a_2 & a_3 \end{pmatrix} \circ \cdots \circ \begin{pmatrix} a_{r-1} & a_r \end{pmatrix}.$$

Since every permutation σ is a product of disjoint cycles by Proposition 3.27, it follows that σ is a product of transpositions. This gives us the desired parity test.

Theorem 3.33. Parity Theorem. Every permutation $\sigma \in S_n$ is a product of transpositions. Moreover, if σ is a product of m transpositions in any way at all, the parity of σ equals the parity of m. That is, σ is even if m is even, and σ is odd if m is odd.

Proof. Write $\sigma = \tau_1 \tau_2 \cdots \tau_m$, where the τ_i are transpositions. If D is the discriminant in n variables, then $\tau_i D = -D$ for each i by Proposition 3.31. Hence the effect of $\sigma = \tau_1 \tau_2 \cdots \tau_m$ on D is to change the sign m times. Thus $\sigma D = (-1)^m D$, and the result follows. □

The following result is now a consequence of Proposition 3.32.

Corollary 3.34. An n-cycle is an even permutation if n is odd and an odd permutation if n is even.

Example 3.35. Write the permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 1 & 8 & 2 & 7 & 3 & 6 & 5 \end{pmatrix}$$

as a product of disjoint cycles and determine its order and parity.

Solution. As disjoint cycles $\pi = \begin{pmatrix} 1 & 4 & 2 \end{pmatrix} \circ \begin{pmatrix} 3 & 8 & 5 & 7 & 6 \end{pmatrix}$. Hence the order of π is lcm $(3, 5) = 15$. The parity of the 3-cycle $\begin{pmatrix} 1 & 4 & 2 \end{pmatrix}$ is even, and the parity of the 5-cycle $\begin{pmatrix} 3 & 8 & 5 & 7 & 6 \end{pmatrix}$ is even; therefore, the parity of π is (even) $\circ$ (even) = even. □

Denote the set of even permutations on n elements by A_n. It follows from Theorem 3.33 that A_n is a subgroup of S_n, called the **alternating group** on n elements. For example,

$$A_4 = \left\{ (1), \begin{array}{l} (1\ 2) \circ (3\ 4), \\ (1\ 3) \circ (2\ 4), \\ (1\ 4) \circ (2\ 3), \end{array} \begin{array}{llll} (1\ 2\ 3), & (1\ 2\ 4), & (1\ 3\ 4), & (2\ 3\ 4), \\ (1\ 3\ 2), & (1\ 4\ 2), & (1\ 4\ 3), & (2\ 4\ 3), \end{array} \right\},$$

a group of 12 elements. In fact, we have

Theorem 3.36. $|A_n| = \frac{1}{2}n!$ for every $n \geqslant 2$.

Proof. Let O_n denote the set of odd permutations in S_n, so that $S_n = A_n \cup O_n$ and $A_n \cap O_n = \emptyset$. Hence $n! = |S_n| = |A_n| + |O_n|$, so it suffices to show that $|A_n| = |O_n|$. We do this by finding a bijection $f: A_n \to O_n$. To this end, write $\tau = \begin{pmatrix} 1 & 2 \end{pmatrix}$ and define f by $f(\sigma) = \tau \circ \sigma$ for all σ in A_n ($\tau \circ \sigma$ is odd because σ is even and τ is odd). Then f is injective because $f(\sigma) = f(\sigma_1)$ implies that $\tau \circ \sigma = \tau \circ \sigma_1$, so $\sigma = \sigma_1$ by cancellation in S_n. To see that f is surjective, let $\lambda \in O_n$. Then $\tau \circ \lambda \in A_n$ and $f(\tau \circ \lambda) = \tau \circ (\tau \circ \lambda) = \lambda$ because $\tau \circ \tau = \varepsilon$. Thus f is surjective, as required. □

Proposition 3.37. Every even permutation can be written as a product of 3-cycles (not necessarily disjoint).

Proof. By Theorem 3.33, an even permutation can be written as a product of an even number of transpositions. We show that any product of two transpositions is a product of 3-cycles. If these two transpositions are identical, their product is the identity. If the two transpositions have one element in common, say (ab) and (bc), their product $(ab) \circ (bc) = (abc)$, a 3-cycle. If the two transpositions have no elements in common, say (ab) and (cd), we can write their product as

$$(ab) \circ (cd) = (ab) \circ (bc) \circ (bc) \circ (cd) = (abc) \circ (bcd),$$

a product of two 3-cycles. □

Theorem 3.33 and Proposition 3.37 show, respectively, that S_n is generated by the 2-cycles and A_n is generated by the 3-cycles.

CAYLEY'S REPRESENTATION THEOREM

At the beginning of the nineteenth century, groups appeared only in a very concrete form, such as symmetry groups or permutation groups. Arthur Cayley (1821–1895) was the first mathematician to deal with groups abstractly in terms of axioms, but he showed that any abstract group can be considered as a subgroup of a symmetric group. Hence, in some sense, if you know all about symmetry groups and permutation groups, you know all about group theory. This result is analogous to Stone's representation theorem for boolean algebras, which proves that any abstract boolean algebra can be considered as an algebra of subsets of a set.

Theorem 3.38. Cayley's Theorem. Every group $(G, \cdot)$ is isomorphic to a subgroup of its symmetric group $(S(G), \circ)$.

Proof. For each element $g \in G$, define $\pi_g: G \to G$ by $\pi_g(x) = g \cdot x$. We show that π_g is a bijection. It is surjective because, for any $y \in G, \pi_g(g^{-1} \cdot y) = g \cdot (g^{-1} \cdot y) = y$. It is injective because $\pi_g(x) = \pi_g(y)$ implies that $g \cdot x = g \cdot y$, and so, by Proposition 3.7, $x = y$. Hence $\pi_g \in S(G)$.

Let $H = \{\pi_g \in S(G) | g \in G\}$. We show that $(H, \circ)$ is a subgroup of $(S(G), \circ)$ isomorphic to $(G, \cdot)$. In fact, we show that the function $\psi: G \to H$ by $\psi(g) = \pi_g$ is a group isomorphism. This is clearly surjective. It is also injective because $\psi(g) = \psi(h)$ implies that $\pi_g = \pi_h$, and $\pi_g(e) = \pi_h(e)$ implies that $g = h$.

It remains to show that ψ preserves the group operation. If $g, h \in G, \pi_{g \cdot h}(x) = (g \cdot h)(x) = g \cdot (h \cdot x) = \pi_g(h \cdot x) = (\pi_g \circ \pi_h)(x)$ and $\pi_{g \cdot h} = \pi_g \circ \pi_h$. Also, $\pi_{h^{-1}} \circ \pi_h = \pi_{h^{-1} \cdot h} = \pi_e$; thus $(\pi_h)^{-1} = \pi_{h^{-1}} \in H$. Hence H is a subgroup of $S(G)$, and $\psi(g \cdot h) = \psi(g) \circ \psi(h)$. □

Corollary 3.39. If G is a finite group of order n, then G is isomorphic to a subgroup of S_n.

Proof. This follows because $S(G)$ is isomorphic to S_n. □

This is not of very much practical value, however, because S_n has order $n!$, which is much larger than the order of G, in general.

EXERCISES

Construct tables for the groups designated in Exercises 3.1 to 3.4.

3.1. C_5.

3.2 D_4.

3.3. $(\mathcal{P}(X), \Delta)$, where $X = \{a, b, c\}$.

3.4 A_4.

For Exercises 3.5 to 3.16, state which are groups and which are abelian groups. Give reasons for your answers.

3.5. $(M_n(\mathbb{R}), +)$.

3.6. $(\{1, 2, 3, 4, 6, 12\},\ \text{gcd})$.

3.7. $(\{a + b\sqrt{2} | a, b \in \mathbb{Q}\}, +)$.

3.8. $(\{a/b \in \mathbb{Q} | a, b \in \mathbb{Z}, b \text{ odd}\}, +)$.

3.9. $(\{z \in \mathbb{C} | |z| = 1\}, +)$.

3.10. $(\{z \in \mathbb{C} | |z| = 1\}, \cdot)$.

3.11. $\left(\left\{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}\right\}, \cdot\right)$.

3.12. $(M_n(\mathbb{R}) - \{0\}, \cdot)$, where 0 is the $n \times n$ zero matrix.

3.13. $(\{1, \zeta, \zeta^2, \zeta^3, \ldots, \zeta^{n-1}\}, \cdot)$, where ζ is a complex nth root of 1.

3.14. $(\{e, a\}, \star)$, where $e \star e = e$ and $e \star a = a \star e = a \star a = a$.

3.15. $(\mathbb{R}^*, \sim)$, where $\mathbb{R}^* = \mathbb{R} - \{0\}$ and $x \sim y$ is xy if $x > 0$, and x/y if $x < 0$.

3.16. $(\mathbb{Z}, \star)$, where $m \star n$ is $m + n$ if m is even, and $m - n$ if m is odd.

3.17. Prove that in any group $(G, \cdot)$, $(a_1 \cdots a_n)^{-1} = a_n^{-1} \cdots a_1^{-1}$.

3.18. If k is an integer (positive or negative), prove that $(a^{-1}ba)^k = a^{-1}b^k a$ in any group $(G, \cdot)$.

3.19. If G is a group in which $a^2 = e$, the identity for all $a \in G$, show that G is abelian.

3.20. Prove that G is abelian if and only if $(ab)^2 = a^2b^2$ for all $a, b \in G$.

3.21. If a is not the identity in a group and $a^4b = ba^5$, prove that $ab \neq ba$.

3.22. Prove that the order of the element g^{-1} is the same as the order of g.

3.23. Prove that the order of the element ab is the same as the order of ba.

3.24. Prove that every image of a cyclic group is cyclic.

3.25. The **gaussian integers** comprise the set of complex numbers $\mathbb{Z}[i] = \{a + ib | a, b \in \mathbb{Z}\}$. Is the group $(\mathbb{Z}[i], +)$ a cyclic group?

For Exercises 3.26 to 3.33, describe the symmetry groups of the figures.

3.26. 

3.27. 

3.28. 

3.29. 

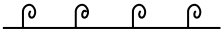
3.30. 

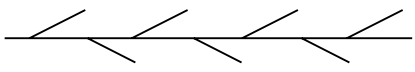
3.31. 

3.32. 

3.33. 

For Exercises 3.34 and 3.35, describe the symmetry groups of the frieze patterns. These patterns are repeated indefinitely in both directions.

3.34. 

3.35. 

3.36. Prove that the relation of being a subgroup is a partial order on the set of subgroups of a group G.

3.37. Draw the poset diagram of the subgroups of C_6.

3.38. Draw the poset diagram of the subgroups of S_3.

3.39. If H and K are subgroups of a group G, prove that $H \cap K$ is also a subgroup of G. Is $H \cup K$ necessarily a subgroup of G?

3.40. If $f: G \to H$ and $g: H \to K$ are group morphisms, show that $g \circ f: G \to K$ is also a group morphism.

3.41. Find all the group morphisms from $(\mathbb{Z}, +)$ to $(\mathbb{Q}, +)$.

3.42. Show that the set $\{f_1, f_2, f_3, f_4, f_5, f_6\}$ of functions $\mathbb{R} - \{0, 1\} \to \mathbb{R} - \{0, 1\}$ under composition is isomorphic to S_3, where

$$f_1(x) = x, \quad f_2(x) = 1 - x, \quad f_3(x) = \frac{1}{x}, \quad f_4(x) = 1 - \frac{1}{x},$$
$$f_5(x) = \frac{1}{1-x}, \quad f_6(x) = \frac{x}{x-1}.$$

3.43. Is $(\mathbb{Z}, +)$ isomorphic to $(\mathbb{Q}^*, \cdot)$, where $\mathbb{Q}^* = \mathbb{Q} - \{0\}$? Give reasons.

3.44. Is $(\mathbb{R}, +)$ isomorphic to $(\mathbb{R}^+, \cdot)$, where $\mathbb{R}^+ = \{x \in \mathbb{R} | x > 0\}$? Give reasons.

3.45. Find the orders of all the elements in A_4.

3.46. Is $A_4 \cong D_6$? Give reasons.

3.47. Draw the table and find the order of all the elements in the group $(\{\pm 1, \pm i, \pm j, \pm k\}, \cdot)$, where $i^2 = j^2 = k^2 = -1$, $ij = k = -ji$, $jk = i = -kj$, and $ki = j = -ik$. This is called the **quaternion group** Q of order 8.

3.48. Let G be the group generated by the matrices $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ under matrix multiplication. Show that G is a non-abelian group of order 8. Is it isomorphic to D_4 or the quaternion group Q?

3.49. Show that D_k is isomorphic to the group generated by $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} \zeta & 0 \\ 0 & \zeta^{-1} \end{pmatrix}$ under matrix multiplication, where $\zeta = \exp(2\pi i/k)$, a complex kth root of unity.

3.50. Construct the table for the group generated by g and h, where g and h satisfy the relations $g^3 = h^2 = e$ and $gh = hg^2$.

3.51. Prove that a group of even order always has at least one element of order 2.

3.52. Find a subgroup of S_7 of order 10.

3.53. Find a subgroup of S_5 of order 3.

3.54. Find a subgroup of A_4 isomorphic to the Klein 4-group.

Multiply out the permutations in Exercises 3.55 to 3.58.

3.55. $\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 3 & 1 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$.

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

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

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

For Exercises 3.59 to 3.62, write the permutations as a product of disjoint cycles. Find the order of each permutation and state whether the permutation is even or odd.

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

3.60. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 4 & 6 & 1 & 5 & 7 & 3 \end{pmatrix}$.

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

3.62. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 8 & 9 & 4 & 2 & 7 & 3 & 5 & 1 & 6 \end{pmatrix}$.

Find the permutations for Exercises 3.63 to 3.66.

3.63. $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 3 & 4 \end{pmatrix}^{-1}$.

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

3.65. $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}^{-1}$.

3.66. $\begin{pmatrix} 1 & 2 & 4 & 6 & 5 & 7 \end{pmatrix}^{-2}$.

For each polynomial in Exercises 3.67 to 3.69, find the permutations of the subscripts that leave the value of the polynomial unchanged. These will form subgroups of S_4, called the symmetry groups of the polynomials.

3.67. $(x_1 + x_2)(x_3 + x_4)$.

3.68. $(x_1 - x_2)(x_3 - x_4)$.

3.69. $(x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^2 + (x_4 - x_1)^2$.

3.70. Describe the group of proper rotations of the tetrahedron with vertices (0, 0, 0), (1, 0, 0), (0, 1, 0), and (0, 0, 1) in $\mathbb{R}^3$.

3.71. Write $G = \{1, -1\}$, a multiplicative subgroup of $\mathbb{R}^+$, and define $f: S_n \to G$ by $f(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd} \end{cases}$. Prove that f is an onto group morphism.

3.72. If g and h are elements in a group G, show that g and $h^{-1}gh$ have the same order.

3.73. What is the number of generators of the cyclic group C_n?

3.74. Express $(123) \circ (456)$ as the power of a single cycle in S_6. Can you generalize this result?

3.75. A perfect interlacing shuffle of a deck of $2n$ cards is the permutation $\begin{pmatrix} 1 & 2 & 3 & \cdots & n & n+1 & n+2 & \cdots & 2n \\ 2 & 4 & 6 & \cdots & 2n & 1 & 3 & \cdots & 2n-1 \end{pmatrix}$. What is the least number of perfect shuffles that have to be performed on a deck of 52 cards before the cards are back in their original position? If there were 50 cards, what would be the least number?

3.76. The **center** of a group G is the set $Z(G) = \{x \in G | xg = gx \text{ for all } g \in G\}$. Show that $Z(G)$ is an abelian subgroup of G.

3.77. Find the center of D_3.

3.78. Find the center of D_4.

3.79. Prove that S_n is generated by the elements (12), (23), (34), ..., $(n-1\,n)$.

3.80. Prove that S_n is generated by the elements $(123\cdots n)$ and (12).

3.81. Prove that A_n is generated by the set $\{(12r)|r = 3, 4, \ldots, n\}$.

3.82. The well-known 15-puzzle consists of a shallow box filled with 16 small squares in a 4×4 array. The bottom right corner square is removed, and the other squares are labeled as in Figure 3.10. By sliding the squares around (without lifting them up), show that the set of possible permutations that can be obtained with the bottom right square blank is precisely A_{15}. (There is no known easy proof that all elements in A_{15} must occur.)

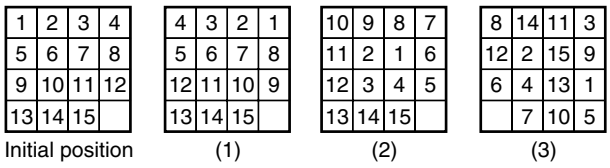


Figure 3.10. The 15-puzzle.

3.83. Which of the positions of the 15-puzzle shown in Figure 3.10 can be achieved?

3.84. An **automorphism** of a group G is an isomorphism from G to itself. Prove that the set of all automorphisms of G forms a group under composition.

3.85. Find the automorphism group of the Klein 4-group.

3.86. Find the automorphism group of C_3.

3.87. Find the automorphism group of C_4.

3.88. Find the automorphism group of S_3.

3.89. A word on $\{x, y\}$ is a finite string of the symbols x, x^{-1}, y, y^{-1}, where x and x^{-1} cannot be adjacent and y and y^{-1} cannot be adjacent; for example, $xxy^{-1}x$ and $x^{-1}x^{-1}yxy$ are words. Let F be the set of such words together with the empty word, which is denoted by 1. The operation of concatenation places one word after another. Show that F is a group under concatenation, where any strings of the form $xx^{-1}, x^{-1}x, yy^{-1}, y^{-1}y$ are deleted in a concatenated word. F is called the **free group on two generators**. Is F abelian? What is the inverse of $x^{-1}x^{-1}yxy$?

4

QUOTIENT GROUPS

Certain techniques are fundamental to the study of algebra. One such technique is the construction of the quotient set of an algebraic object by means of an equivalence relation on the underlying set. For example, if the object is the group of integers $(\mathbb{Z}, +)$, the congruence relation modulo n on $\mathbb{Z}$ will define the quotient group of integers modulo n.

This quotient construction can be applied to numerous algebraic structures, including groups, boolean algebras, and vector spaces.

In this chapter we introduce the concept of an equivalence relation and go on to apply this to groups. We obtain Lagrange's theorem, which states that the order of a subgroup divides the order of the group, and we also obtain the morphism theorem for groups. We study the implications of these two theorems and classify the groups of low order.

EQUIVALENCE RELATIONS

Relations are one of the basic building blocks of mathematics (as well as of the rest of the world). A **relation** R from a set S to a set T is a subset of $S \times T$. We say that a is related to b under R if the pair (a, b) belongs to the subset, and we write this as aRb. If (a, b) does not belong to the subset, we say that a is not related to b, and write $a \not R b$. This definition even covers many relations in everyday life, such as "is the father of," "is richer than," and "goes to the same school as" as well as mathematical relations such as "is equal to," "is a member of," and "is similar to." A relation R from S to T has the property that for any elements a in S, and b in T, either aRb or $a \not R b$.

Any function $f: S \to T$ gives rise to a relation R from S to T by taking aRb to mean $f(a) = b$. The subset R of $S \times T$ is the graph of the function. However, relations are much more general than functions. One element can be related to many elements or to no elements at all.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

A relation from a set S to itself is called a **relation on** S. Any partial order on a set, such as "$\leqslant$" on the real numbers, or "is a subset of" on a power set $\mathcal{P}(X)$, is a relation on that set. "Equals" is a relation on any set S and is defined by the subset $\{(a, a)|a \in S\}$ of $S \times S$. An equivalence relation is a relation that has the most important properties of the "equals" relation.

A relation E on a set S is called an **equivalence relation** if the following conditions hold.

(i) aEa for all $a \in S$.	(reflexive condition)
(ii) If aEb, then bEa.	(symmetric condition)
(iii) If aEb and bEc, then aEc.	(transitive condition)

If E is an equivalence relation on S and $a \in S$, then $[a] = \{x \in S|xEa\}$ is called the **equivalence class** containing a. The set of all equivalence classes is called the **quotient set** of S by E and is denoted by S/E. Hence

$$S/E = \{[a]|a \in S\}.$$

Proposition 4.1. If E is an equivalence relation on a set S, then

(i) If aEb, then $[a] = [b]$.
(ii) If $a \not E b$, then $[a] \cap [b] = \emptyset$.
(iii) S is the disjoint union of all the distinct equivalence classes.

Proof. (i) If aEb, let x be any element of $[a]$. Then xEa and so xEb by transitivity. Hence $x \in [b]$ and $[a] \subseteq [b]$. The symmetry of E implies that bEa, and an argument similar to the above shows that $[b] \subseteq [a]$. This proves that $[a] = [b]$.

(ii) Suppose that $a \not E b$. If there was an element $x \in [a] \cap [b]$, then xEa, xEb, so aEb by symmetry and transitivity. Hence $[a] \cap [b] = \emptyset$.

(iii) Parts (i) and (ii) show that two equivalence classes are either the same or disjoint. The reflexivity of E implies that each element $a \in S$ is in the equivalence class $[a]$. Hence S is the disjoint union of all the equivalence classes. □

A collection of nonempty subsets is said to **partition** a set S if the union of the subsets is S and any two subsets are disjoint. The previous proposition shows that any equivalence relation partitions the set into its equivalence classes. Each element of the set belongs to one and only one equivalence class.

It can also be shown that every partition of a set gives rise to an equivalence relation whose classes are precisely the subsets in the partition.

Example 4.2. Let n be a fixed positive integer and a and b any two integers. We say that a **is congruent to** b **modulo** n if n divides $a - b$. We denote this by

$$a \equiv b \bmod n.$$

Show that this congruence relation modulo n is an equivalence relation on $\mathbb{Z}$. The set of equivalence classes is called the set of **integers modulo** n and is denoted by $\mathbb{Z}_n$.

Solution. Write "$n|m$" for "n divides m," which means that there is some integer k such that $m = nk$. Hence $a \equiv b \bmod n$ if and only if $n|(a - b)$.

(i) For all $a \in \mathbb{Z}$, $n|(a - a)$, so $a \equiv a \bmod n$ and the relation is reflexive.
(ii) If $a \equiv b \bmod n$, then $n|(a - b)$, so $n|-(a - b)$. Hence $n|(b - a)$ and $b \equiv a \bmod n$.
(iii) If $a \equiv b \bmod n$ and $b \equiv c \bmod n$, then $n|(a - b)$ and $n|(b - c)$, so $n|(a - b) + (b - c)$. Therefore, $n|(a - c)$ and $a \equiv c \bmod n$.

Hence congruence modulo n is an equivalence relation on $\mathbb{Z}$. □

In the congruence relation modulo 3, we have the following equivalence classes:

$$[0] = \{\ldots, -3, 0, 3, 6, 9, \ldots\}$$
$$[1] = \{\ldots, -2, 1, 4, 7, 10, \ldots\}$$
$$[2] = \{\ldots, -1, 2, 5, 8, 11, \ldots\}$$
$$[3] = \{\ldots, 0, 3, 6, 9, 12, \ldots\} = [0]$$

Any equivalence class must be one of [0], [1], or [2], so $\mathbb{Z}_3 = \{[0], [1], [2]\}$.

In general, $\mathbb{Z}_n = \{[0], [1], [2], \ldots, [n - 1]\}$, since any integer is congruent modulo n to its remainder when divided by n.

One set of equivalence classes that is introduced in elementary school is the set of rational numbers. Students soon become used to the fact that $\frac{1}{2}$ and $\frac{3}{6}$ represent the same rational number. We need to use the concept of equivalence class to define a rational number precisely. Define the relation E on $\mathbb{Z} \times \mathbb{Z}^*$ (where $\mathbb{Z}^* = \mathbb{Z} - \{0\}$) by $(a, b)\ E\ (c, d)$ if and only if $ad = bc$. This is an equivalence relation on $\mathbb{Z} \times \mathbb{Z}^*$, and the equivalence classes are called **rational numbers**. We denote the equivalence class $[(a, b)]$ by $\frac{a}{b}$. Therefore, since $(1, 2)\ E(3, 6)$, it follows that $\frac{1}{2} = \frac{3}{6}$.

Two series-parallel circuits involving the switches $A_1, A_2, \ldots, A_n$ are said to be *equivalent* if they both are open or both are closed for any position of the n switches. This is an equivalence relation, and the equivalence classes are the 2^{2^n} distinct types of circuits controlled by n switches.

Any permutation π on a set S induces an equivalence relation, $\sim$, on S where $a \sim b$ if and only if $b = \pi^r(a)$, for some $r \in \mathbb{Z}$. The equivalence classes are the **orbits** of π. In the decomposition of the permutation π into disjoint cycles, the elements in each cycle constitute one orbit.

COSETS AND LAGRANGE'S THEOREM

The congruence relation modulo n on $\mathbb{Z}$ can be defined by $a \equiv b \bmod n$ if and only if $a - b \in n\mathbb{Z}$, where $n\mathbb{Z}$ is the subgroup of $\mathbb{Z}$ consisting of all multiples

of n. We now generalize this notion and define congruence in any group modulo one of its subgroups. We are interested in the equivalence classes, which we call *cosets*.

Let $(G, \cdot)$ be a group with subgroup H. For $a, b \in G$, we say that a **is congruent to** b **modulo** H, and write $a \equiv b \bmod H$ if and only if $ab^{-1} \in H$.

Proposition 4.3. The relation $a \equiv b \bmod H$ is an equivalence relation on G. The equivalence class containing a can be written in the form $Ha = \{ha | h \in H\}$, and it is called a **right coset** of H in G. The element a is called a **representative** of the coset Ha.

Proof. (i) For all $a \in G$, $aa^{-1} = e \in H$; thus the relation is reflexive.

(ii) If $a \equiv b \bmod H$, then $ab^{-1} \in H$; thus $ba^{-1} = (ab^{-1})^{-1} \in H$. Hence $b \equiv a \bmod H$, and the relation is symmetric.

(iii) If $a \equiv b$ and $b \equiv c \bmod H$, then ab^{-1} and $bc^{-1} \in H$. Hence $ac^1 = (ab^{-1})(bc^{-1}) \in H$ and $a \equiv c \bmod H$. The relation is transitive. Hence $\equiv$ is an equivalence relation. The equivalence class containing a is

$$\begin{aligned} \{x \in G | x \equiv a \bmod H\} &= \{x \in G | xa^{-1} = h \in H\} \\ &= \{x \in G | x = ha, \text{ where } h \in H\} \\ &= \{ha | h \in H\}, \end{aligned}$$

which we denote by Ha. □

Example 4.4. Find the right cosets of A_3 in S_3.

Solution. One coset is the subgroup itself $A_3 = \{(1), (123), (132)\}$. Take any element not in the subgroup, say (12). Then another coset is

$$A_3(12) = \{(12), (123) \circ (12), (132) \circ (12)\} = \{(12), (13), (23)\}.$$

Since the right cosets form a partition of S_3 and the two cosets above contain all the elements of S_3, it follows that these are the only two cosets.

In fact, $A_3 = A_3(123) = A_3(132)$ and $A_3(12) = A_3(13) = A_3(23)$. □

Example 4.5. Find the right cosets of $H = \{e, g^4, g^8\}$ in $C_{12} = \{e, g, g^2, \ldots, g^{11}\}$.

Solution. H itself is one coset. Another is $Hg = \{g, g^5, g^9\}$. These two cosets have not exhausted all the elements of C_{12}, so pick an element, say g^2, which is not in H or Hg. A third coset is $Hg^2 = \{g^2, g^6, g^{10}\}$ and a fourth is $Hg^3 = \{g^3, g^7, g^{11}\}$.

Since $C_{12} = H \cup Hg \cup Hg^2 \cup Hg^3$, these are all the cosets. □

As the examples above suggest, every coset contains the same number of elements. We use this result to prove the famous theorem of Joseph Lagrange (1736–1813).

Lemma 4.6. There is a bijection between any two right cosets of H in G.

Proof. Let Ha be a right coset of H in G. We produce a bijection between Ha and H, from which it follows that there is a bijection between any two right cosets.

Define $\psi: H \to Ha$ by $\psi(h) = ha$. Then ψ is clearly surjective. Now suppose that $\psi(h_1) = \psi(h_2)$, so that $h_1a = h_2a$. Multiplying each side by a^{-1} on the right, we obtain $h_1 = h_2$. Hence ψ is a bijection. □

Theorem 4.7. Lagrange's Theorem. If G is a finite group and H is a subgroup of G, then $|H|$ divides $|G|$.

Proof. The right cosets of H in G form a partition of G, so G can be written as a disjoint union

$$G = Ha_1 \cup Ha_2 \cup \cdots \cup Ha_k \text{ for a finite set of elements } a_1, a_2, \ldots, a_k \in G.$$

By Lemma 4.6, the number of elements in each coset is $|H|$. Hence, counting all the elements in the disjoint union above, we see that $|G| = k|H|$. Therefore, $|H|$ divides $|G|$. □

If H is a subgroup of G, the number of distinct right cosets of H in G is called the **index** of H in G and is written $|G : H|$. The following is a direct consequence of the proof of Lagrange's theorem.

Corollary 4.8. If G is a finite group with subgroup H, then

$$|G : H| = |G|/|H|.$$

Corollary 4.9. If a is an element of a finite group G, then the order of a divides the order of G.

Proof. Let $H = \{a^r | r \in \mathbb{Z}\}$ be the cyclic subgroup generated by a. By Proposition 3.13, the order of the subgroup H is the same as the order of the element a. Hence, by Lagrange's theorem, the order of a divides the order of G. □

Corollary 4.10. If a is an element of the finite group G, then $a^{|G|} = e$.

Proof. If m is the order of a, then $|G| = mk$ for some integer k. Hence $a^{|G|} = a^{mk} = (a^m)^k = e^k = e$. □

Corollary 4.11. If G is a group of prime order, then G is cyclic.

Proof. Let $|G| = p$, a prime number. By Corollary 4.9, every element has order 1 or p. But the only element of order 1 is the identity. Therefore, all the

other elements have order p, and there is at least one because $|G| \geqslant 2$. Hence by Theorem 3.14, G is a cyclic group. □

The converse of Lagrange's theorem is false, as the following example shows. That is, if k is a divisor of the order of G, it does not necessarily follow that G has a subgroup of order k.

Example 4.12. A_4 is a group of order 12 having no subgroup of order 6.

Solution. A_4 contains one identity element, eight 3-cycles of the form (abc), and three pairs of transpositions of the form $(ab) \circ (cd)$, where a, b, c, and d are distinct elements of $\{1, 2, 3, 4\}$. If a subgroup contains a 3-cycle (abc), it must also contain its inverse (acb). If a subgroup of order 6 exists, it must contain the identity and a product of two transpositions, because the odd number of nonidentity elements cannot be made up of 3-cycles and their inverses. A subgroup of order 6 must also contain at least two 3-cycles because A_4 only contains four elements that are not 3-cycles.

Without loss of generality, suppose that a subgroup of order 6 contains the elements (abc) and $(ab) \circ (cd)$. Then it must also contain the elements $(abc)^{-1} = (acb)$, $(abc) \circ (ab) \circ (cd) = (acd)$, $(ab) \circ (cd) \circ (abc) = (bdc)$, and $(acd)^{-1} = (adc)$, which, together with the identity, gives more than six elements. Hence A_4 contains no subgroup of order 6. □

The next proposition strengthens Lagrange's theorem in the case of finite cyclic groups. The following lemma, of interest in its own right, will be needed.

Lemma 4.13. Let g be an element of order n in a group, and let $m \geqslant 1$.

(i) If $\gcd(n, m) = d$, then g^m has order n/d.

(ii) In particular, if m divides n, then g^m has order n/m.

Proof. (i). We have $(g^m)^{n/d} = (g^n)^{m/d} = e^{m/d} = e$. If $(g^m)^k = e$, we must show that $\frac{n}{d}$ divides k. We have $g^{mk} = e$, so n divides mk by Proposition 3.11. Hence $\frac{n}{d}$ divides $\frac{m}{d}k$. But $\frac{n}{d}$ and $\frac{m}{d}$ are relatively prime by Theorem 11, Appendix 2, so $\frac{n}{d}$ divides k (by the same theorem).

(ii). If m divides n, then $\gcd(n, m) = m$, so (i) implies (ii). □

Proposition 4.14. If G is a cyclic group of order n, and if k divides n, then G has exactly one subgroup H of order k. In fact, if g generates G, then H is generated by $g^{n/k}$.

Proof. Let H denote the subgroup generated by $g^{n/k}$. Then $|H| = k$ because $g^{n/k}$ has order k by Lemma 4.13 $\left(\text{with } m = \frac{n}{k}\right)$. Now let K be any subgroup of

G of order k. By Proposition 3.12, K is generated by g^m for some $m \in \mathbb{Z}$. Then g^m has order $|K| = k$ by Proposition 3.13. But if $d = \gcd(m, n)$, then g^m also has order n/d by Lemma 4.13. Thus $k = n/d$, so $d = n/k$. Write $d = xm + yn$, x, $y \in \mathbb{Z}$ (by Theorem 8, Appendix 2). Then $g^{n/k} = g^d = (g^m)^x (g^n)^y = (g^m)^x \in K$. Since $g^{n/k}$ generates H, it follows that $H \subseteq K$, so $H = K$ because $|H| = |K|$. □

NORMAL SUBGROUPS AND QUOTIENT GROUPS

Let G be a group with subgroup H. The right cosets of H in G are equivalence classes under the relation $a \equiv b \bmod H$, defined by $ab^{-1} \in H$. We can also define the relation L on G so that aLb if and only if $b^{-1}a \in H$. This relation, L, is an equivalence relation, and the equivalence class containing a is the **left coset** $aH = \{ah | h \in H\}$. As the following example shows, the left coset of an element does not necessarily equal the right coset.

Example 4.15. Find the left and right cosets of $H = A_3$ and $K = \{(1), (12)\}$ in S_3.

Solution. We calculated the right cosets of $H = A_3$ in Example 4.4.

	Right Cosets		Left Cosets
H	$= \{(1), (123), (132)\}$	H	$= \{(1), (123), (132)\}$
$H(12)$	$= \{(12), (13), (23)\}$	$(12)H$	$= \{(12), (23), (13)\}$

In this case, the left and right cosets of H are the same.

However, the left and right cosets of K are not all the same.

	Right Cosets		Left Cosets
K	$= \{(1), (12)\}$	K	$= \{(1), (12)\}$
$K(13)$	$= \{(13), (132)\}$	$(13)K$	$= \{(13), (123)\}$
$K(23)$	$= \{(23), (123)\}$	$(23)K$	$= \{(23), (132)\}$

□

Since $a \equiv b \bmod H$ is an equivalence relation for any subgroup H of a group G and the quotient set is the set of right cosets $\{Ha | a \in G\}$, it is natural to ask whether this quotient *set* is also a *group* with a multiplication induced by the multiplication in G. We show that this is the case if and only if the right cosets of H equal the left cosets.

A subgroup H of a group G is called a **normal subgroup** of G if $g^{-1}hg \in H$ for all $g \in G$ and $h \in H$.

Proposition 4.16. $Hg = gH$, for all $g \in G$, if and only if H is a normal subgroup of G.

Proof. Suppose that $Hg = gH$. Then, for any element $h \in H$, $hg \in Hg = gH$. Hence $hg = gh_1$ for some $h_1 \in H$ and $g^{-1}hg = g^{-1}gh_1 = h_1 \in H$. Therefore, H is a normal subgroup.

Conversely, if H is normal, let $hg \in Hg$ and $g^{-1}hg = h_1 \in H$. Then $hg = gh_1 \in gH$ and $Hg \subseteq gH$. Also, $ghg^{-1} = (g^{-1})^{-1}hg^{-1} = h_2 \in H$, since H is normal, so $gh = h_2g \in Hg$. Hence, $gH \subseteq Hg$, and so $Hg = gH$. □

Therefore, A_3 is a normal subgroup of S_3 by Example 4.13, whereas {(1), (12)} is not.

Proposition 4.17. Any subgroup of an abelian group is normal.

Proof. If H is a subgroup of an abelian group, G, then $g^{-1}hg = hg^{-1}g = h \in H$ for all $g \in G$, $h \in H$. Hence H is normal. □

If N is a normal subgroup of a group G, the left cosets of N in G are the same as the right cosets of N in G, so there will be no ambiguity in just talking about the cosets of N in G.

Theorem 4.18. If N is a normal subgroup of $(G, \cdot)$, the set of cosets $G/N = \{Ng | g \in G\}$ forms a group $(G/N, \cdot)$, where the operation is defined by $(Ng_1) \cdot (Ng_2) = N(g_1 \cdot g_2)$. This group is called the **quotient group** or **factor group** of G by N.

Proof. The operation of multiplying two cosets, Ng_1 and Ng_2, is defined in terms of particular elements, g_1 and g_2, of the cosets. For this operation to make sense, we have to verify that, if we choose different elements, h_1 and h_2, in the same cosets, the product coset $N(h_1 \cdot h_2)$ is the same as $N(g_1 \cdot g_2)$. In other words, we have to show that multiplication of cosets is well defined.

Since h_1 is in the same coset as g_1, we have $h_1 \equiv g_1 \bmod N$. Similarly, $h_2 \equiv g_2 \bmod N$. We show that $Nh_1h_2 = Ng_1g_2$. We have $h_1g_1^{-1} = n_1 \in N$ and $h_2g_2^{-1} = n_2 \in N$, so $h_1h_2(g_1g_2)^{-1} = h_1h_2g_2^{-1}g_1^{-1} = n_1g_1n_2g_2g_2^{-1}g_1^{-1} = n_1g_1n_2g_1^{-1}$. Now N is a normal subgroup, so $g_1n_2g_1^{-1} \in N$ and $n_1g_1n_2g_1^{-1} \in N$. Hence $h_1h_2 \equiv g_1g_2 \bmod N$ and $Nh_1h_2 = Ng_1g_2$. Therefore, the operation is well defined.

The operation is associative because $(Ng_1 \cdot Ng_2) \cdot Ng_3 = N(g_1g_2) \cdot Ng_3 = N(g_1g_2)g_3$ and also $Ng_1 \cdot (Ng_2 \cdot Ng_3) = Ng_1 \cdot N(g_2g_3) = Ng_1(g_2g_3) = N(g_1g_2)g_3$.

Since $Ng \cdot Ne = Nge = Ng$ and $Ne \cdot Ng = Ng$, the identity is $Ne = N$. The inverse of Ng is Ng^{-1} because $Ng \cdot Ng^{-1} = N(g \cdot g^{-1}) = Ne = N$ and also $Ng^{-1} \cdot Ng = N$.

Hence $(G/N, \cdot)$ is a group. □

The order of G/N is the number of cosets of N in G. Hence

$$|G/N| = |G : N| = |G|/|N|.$$

TABLE 4.1. Quotient Group S_3/A_3

$\circ$	H	$H(12)$
H	H	$H(12)$
$H(12)$	$H(12)$	H

We have seen in Example 4.15 that A_3 is a normal subgroup of S_3; therefore, S_3/A_3 is a quotient group. If $H = A_3$, the elements of this group are the cosets H and $H(12)$, and its multiplication table is given in Table 4.1.

Example 4.19. $(\mathbb{Z}_n, +)$ is the quotient group of $(\mathbb{Z}, +)$ by the subgroup $n\mathbb{Z} = \{nz | z \in \mathbb{Z}\}$.

Solution. Since $(\mathbb{Z}, +)$ is abelian, every subgroup is normal. The set $n\mathbb{Z}$ can be verified to be a subgroup, and the relationship $a \equiv b \bmod n\mathbb{Z}$ is equivalent to $a - b \in n\mathbb{Z}$ and to $n|a - b$. Hence $a \equiv b \bmod n\mathbb{Z}$ is the same relation as $a \equiv b \bmod n$. Therefore, $\mathbb{Z}_n$ is the quotient group $\mathbb{Z}/n\mathbb{Z}$, where the operation on congruence classes is defined by $[a] + [b] = [a + b]$. □

$(\mathbb{Z}_n, +)$ is a cyclic group with 1 as a generator, and therefore, by Theorem 3.25, is isomorphic to C_n. The group $(\mathbb{Z}_5, +)$ is shown in Table 4.2.

When there is no confusion, we write the elements of $\mathbb{Z}_n$ as $0, 1, 2, 3, \ldots, n - 1$ instead of $[0], [1], [2], [3], \ldots, [n - 1]$.

Proposition 4.20. If H is a subgroup of index 2 in G, so that $|G : H| = 2$, then H is a normal subgroup of G, and G/H is cyclic group of order 2.

Proof. Since $|G : H| = 2$, there are only two right cosets of H in G. One must be H and the other can be written as Hg, where g is any element of G that is not in H. To show that H is a normal subgroup of G, we need to show that $g^{-1}hg \in H$ for all $g \in G$ and $h \in H$. If g is an element of H, it is clear that $g^{-1}hg \in H$ for all $h \in H$. If g is not an element of H, suppose that $g^{-1}hg \notin H$. In this case, $g^{-1}hg$ must be an element of the other right coset Hg, and we can write $g^{-1}hg = h_1 g$, for some $h_1 \in H$. It follows that $g = hh_1^{-1} \in H$, which contradicts the fact that $g \notin H$. Hence $g^{-1}hg \in H$ for all $g \in G$ and $h \in H$; in other words, H is normal in G. □

TABLE 4.2. Group $(\mathbb{Z}_5, +)$

$+$	[0]	[1]	[2]	[3]	[4]
[0]	[0]	[1]	[2]	[3]	[4]
[1]	[1]	[2]	[3]	[4]	[0]
[2]	[2]	[3]	[4]	[0]	[1]
[3]	[3]	[4]	[0]	[1]	[2]
[4]	[4]	[0]	[1]	[2]	[3]

Theorem 4.21. If G is a finite abelian group and the prime p divides the order of G, then G contains an element of order p and hence a subgroup of order p.

Proof. We prove this result by induction on the order of G. For a particular prime p, suppose that all abelian groups of order less than k, whose order is divisible by p, contain an element of order p. The result is vacuously true for groups of order 1. Now suppose that G is a group of order k. If p divides k, choose any nonidentity element $g \in G$. Let t be the order of the element g.

Case 1. If p divides t, say $t = pr$, then g^r is an element of order p. This follows because g^r is not the identity, but $(g^r)^p = g^t = e$, and p is a prime.

Case 2. On the other hand, if p does not divide t, let K be the subgroup generated by g. Since G is abelian, K is normal, and the quotient group G/K has order $|G|/t$, which is divisible by p. Therefore, by the induction hypothesis, G/K has an element of order p, say Kh. If u is the order of h in G, then $h^u = e$ and $(Kh)^u = Kh^u = K$. Since Kh has order p in G/K, u is a multiple of p, and we are back to case 1.

The result now follows from the principle of mathematical induction. □

This result is a partial converse to Lagrange's theorem. It is a special case of some important results, in more advanced group theory, known as the Sylow theorems. These theorems give information on the subgroups of prime power order, and they can be found in books such as Herstein [9], Hall [30], or Nicholson [11].

Example 4.22. Show that A_5 has no proper normal subgroups.

Solution. It follows from Corollary 3.34 that A_5 contains three types of nonidentity elements: 3-cycles, 5-cycles, and pairs of disjoint transpositions. Suppose that N is a normal subgroup of A_5 that contains more than one element.

Case 1. Suppose that N contains the 3-cycle (abc). From the definition of normal subgroup, $g^{-1} \circ (abc) \circ g \in N$ for all $g \in A_5$. If we take $g = (ab) \circ (cd)$, we obtain

$$(ab) \circ (cd) \circ (abc) \circ (ab) \circ (cd) = (adb) \in N$$

and also $(adb)^{-1} = (abd) \in N$. In a similar way, we can show that N contains every 3-cycle. Therefore, by Proposition 3.37, N must be the entire alternating group.

Case 2. Suppose that N contains the 5-cycle $(abcde)$. Then

$$\text{and } (abc)^{-1} \circ (abcde) \circ (abc) = (acb) \circ (abcde) \circ (abc) = (abdec) \in N$$

$$(abcde) \circ (abdec)^{-1} = (abcde) \circ (acedb) = (adc) \in N.$$

We are now back to case 1, and hence $N = A_5$.

Case 3. Suppose that N contains the pair of disjoint transpositions $(ab) \circ (cd)$. Then, if e is the element of $\{1, 2, 3, 4, 5\}$ not appearing in these transpositions, we have

$$(abe)^{-1} \circ (ab) \circ (cd) \circ (abe) = (aeb) \circ (ab) \circ (cd) \circ (abe) = (ae) \circ (cd) \in N.$$

Also, $(ab) \circ (cd) \circ (ae) \circ (cd) = (aeb) \in N$, and again we are back to case 1.

We have shown that any normal subgroup of A_5 containing more than one element must be A_5 itself. □

A group without any proper normal subgroups is called a **simple group**. The term *simple* must be understood in the technical sense that it cannot be broken down, because it cannot have any nontrivial quotient groups. This is analogous to a prime number, which has no nontrivial quotients. Apart from the cyclic groups of prime order, which have no proper subgroups of any kind, simple groups are comparatively rare.

The group A_5 is of great interest to mathematicians because it is used in Galois theory to show that there is an equation of the fifth degree that cannot be solved by any algebraic formula.

It can be shown that every alternating group A_n, $n \geqslant 5$, is simple. The cyclic groups C_p, p a prime, are another infinite series of simple groups (the abelian ones), and other series have been known for decades. But it was not until 1981 that the finite simple groups were completely classified. This was the culmination of more than 30 years of effort by hundreds of mathematicians, yielding thousands of pages of published work, and was one of the great achievements of twentieth-century mathematics. One spectacular landmark came in 1963, when J. G. Thompson and W. Feit verified a long-standing conjecture of W. Burnside (1852–1927) that every finite, non-abelian, simple group has even order (the proof is more than 250 pages long!). The main difficulty in the classification was the existence of *sporadic* finite simple groups, not belonging to any of the known families. The largest of these, known as the *monster*, has order approximately 2×10^{53}. The complete classification encompasses several infinite families and exactly 26 sporadic groups.

MORPHISM THEOREM

The morphism theorem is a basic result of group theory that describes the relationship between morphisms, normal subgroups, and quotient groups. There is an analogous result for most algebraic systems, including rings and vector spaces.

If $f: G \to H$ is a group morphism, the **kernel** of f, denoted by $\mathrm{Ker} f$, is defined to be the set of elements of G that are mapped by f to the identity of H. That is, $\mathrm{Ker} f = \{g \in G | f(g) = e_H\}$.

Proposition 4.23. Let $f: G \to H$ be a group morphism. Then:

(i) $\mathrm{Ker} f$ is a normal subgroup of G.
(ii) f is injective if and only if $\mathrm{Ker} f = \{e_G\}$.

Proof. (i) We first show that $\mathrm{Ker} f$ is a subgroup of G. Let $a, b \in \mathrm{Ker} f$ so that $f(a) = f(b) = e_H$. Then

$$f(ab) = f(a)f(b) = e_H e_H = e_H, \quad \text{so} \quad ab \in \mathrm{Ker} f$$

and

$$f(a^{-1}) = f(a)^{-1} = e_H^{-1} = e_H, \quad \text{so} \quad a^{-1} \in \mathrm{Ker} f.$$

Therefore, $\mathrm{Ker} f$ is a subgroup of G.

If $a \in \mathrm{Ker} f$ and $g \in G$, then

$$f(g^{-1}ag) = f(g^{-1})f(a)f(g) = f(g)^{-1}e_H f(g) = f(g)^{-1}f(g) = e_H.$$

Hence $g^{-1}ag \in \mathrm{Ker} f$, and $\mathrm{Ker} f$ is a normal subgroup of G.

(ii) If f is injective, only one element maps to the identity of **H**. Hence $\mathrm{Ker} f = \{e_G\}$. Conversely, if $\mathrm{Ker} f = \{e_G\}$, suppose that $f(g_1) = f(g_2)$. Then $f(g_1 g_2^{-1}) = f(g_1)f(g_2)^{-1} = e_H$ so $g_1 g_2^{-1} \in \mathrm{Ker} f = \{e_G\}$. Hence $g_1 = g_2$, and f is injective. □

Proposition 4.24. For any group morphism $f: G \to H$, the **image** of f, $\mathrm{Im} f = \{f(g) | g \in G\}$, is a subgroup of H (although not necessarily normal).

Proof. Let $f(g_1), \ f(g_2) \in \mathrm{Im} f$. Then $e_H = f(e_G) \in \mathrm{Im} f$, $f(g_1)f(g_2) = f(g_1 g_2) \in \mathrm{Im} f$, and $f(g_1)^{-1} = f(g_1^{-1}) \in \mathrm{Im} f$. Hence $\mathrm{Im} f$ is a subgroup of H. □

Theorem 4.25. Morphism Theorem for Groups. Let K be the kernel of the group morphism $f: G \to H$. Then G/K is isomorphic to the image of f, and the isomorphism $\psi: G/K \to \mathrm{Im} f$ is defined by $\psi(Kg) = f(g)$.

This result is also known as the **first isomorphism theorem**; the second and third isomorphism theorems are given in Exercises 4.43 and 4.44.

Proof. The function ψ is defined on a coset by using one particular element in the coset, so we have to check that ψ is well defined; that is, it does not matter which element we use. If $Kg = Kg'$, then $g' \equiv g \bmod K$ so $g'g^{-1} = k \in K = \mathrm{Ker} f$. Hence $g' = kg$ and so

$$f(g') = f(kg) = f(k)f(g) = e_H f(g) = f(g).$$

Thus ψ is well defined on cosets.

The function ψ is a morphism because

$$\psi(Kg_1 Kg_2) = \psi(Kg_1 g_2) = f(g_1 g_2) = f(g_1)f(g_2) = \psi(Kg_1)\psi(Kg_2).$$

If $\psi(Kg) = e_H$, then $f(g) = e_H$ and $g \in K$. Hence the only element in the kernel of ψ is the identity coset K, and ψ is injective. Finally, $\mathrm{Im}\, \psi = \mathrm{Im} f$,

by the definition of ψ. Therefore, ψ is the required isomorphism between G/K and $\mathrm{Im} f$. □

Conversely, note that if K is any normal subgroup of G, the map $g \mapsto Kg$ is a morphism from G to G/K, whose kernel is precisely K.

By taking f to be the identity morphism from G to itself, the morphism theorem implies that $G/\{e\} \cong G$.

The function $f: \mathbb{Z} \to \mathbb{Z}_n$, defined by $f(x) = [x]$, has $n\mathbb{Z}$ as its kernel, and therefore the morphism theorem yields the fact that $\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$.

If a and b are generators of the cyclic groups $C_{12} = \langle a \rangle$ and $C_6 = \langle b \rangle$, respectively, consider the function $f: C_{12} \to C_6$ given by $f(a^r) = b^{2r}$. This is well defined. In fact, if $a^r = a^{r_1}$, then 12 divides $r - r_1$ by Proposition 3.11, so certainly $b^{2r} = b^{2r_1}$. It is easily verified that f is morphism, and the kernel is $K = \{e, a^3, a^6, a^9\}$ because if a^r is in K, then $b^{2r} = 0$, so 6 divides $2r$, whence 3 divides r. Thus $C_{12}/K \cong C_6$, and this isomorphism is obtained by mapping the coset Ka^r to b^{2r}.

The alternating group A_n is of index 2 in the symmetric group S_n (by Theorem 3.36) and so is a normal subgroup by Proposition 4.20. It is instructive to obtain this same conclusion from the morphism theorem. If σ is a permutation in S_n, recall that σ is called *even* or *odd* according as $\sigma D = D$ or $\sigma D = -D$, where D is the *discriminant* in n variables (see the discussion leading to Theorem 3.33). Consider the multiplicative group $\{1, -1\}$, and define a function $f: S_n \to \{1, -1\}$ by $f(\sigma) = \begin{cases} 1 & \text{if } \sigma D = D \\ -1 & \text{if } \sigma D = -D \end{cases}$. Then f is a surjective morphism (verify) and the kernel is the group A_n of even permutations. Since $|S_n| = n!$, the morphism theorem and Corollary 4.8 give the following result (and reprove Theorem 3.36).

Proposition 4.26. A_n is a normal subgroup of S_n, $S_n/A_n \cong C_2$, and $|A_n| = \frac{1}{2}n!$.

Example 4.27. Show that the quotient group $\mathbb{R}/\mathbb{Z}$, of real numbers modulo 1 is isomorphic to the **circle group** $W = \{e^{i\theta} \in \mathbb{C} | \theta \in \mathbb{R}\}$.

Solution. The set W consists of points on the circle of complex numbers of unit modulus, and forms a group under multiplication. Define the function $f: \mathbb{R} \to W$ by $f(x) = e^{2\pi i x}$. This is a morphism from $(\mathbb{R}, +)$ to $(W, \cdot)$ because

$$f(x + y) = e^{2\pi i(x+y)} = e^{2\pi i x} \cdot e^{2\pi i y} = f(x) \cdot f(y).$$

This function can be visualized in Figure 4.1 as wrapping the real line around and around the circle.

The morphism f is clearly surjective, and its kernel is $\{x \in \mathbb{R} | e^{2\pi i x} = 1\} = \mathbb{Z}$. Therefore, the morphism theorem implies that $\mathbb{R}/\mathbb{Z} \cong W$. The quotient space $\mathbb{R}/\mathbb{Z}$ is the set of equivalence classes of $\mathbb{R}$ under the relation defined by $x \equiv y \bmod \mathbb{Z}$ if and only if the real numbers x and y differ by an integer. This quotient space is called the *group of real numbers modulo* 1. □

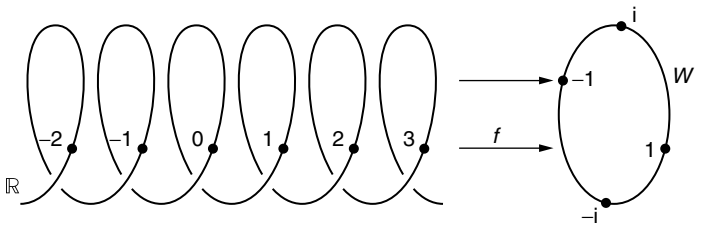


Figure 4.1. Morphism $f: \mathbb{R} \to W$.

Proposition 4.28. If G and H are finite groups whose orders are relatively prime, there is only one morphism from G to H, the trivial one.

Proof. Let K be the kernel of a morphism f from G to H. Then $G/K \cong \text{Im} f$, a subgroup of H. Now $|G/K| = |G|/|K|$, which is a divisor of $|G|$. But by Lagrange's theorem, $|\text{Im} f|$ is a divisor of $|H|$. Since $|G|$ and $|H|$ are relatively prime, we must have $|G/K| = |\text{Im} f| = 1$. Therefore $K = G$, so $f: G \to H$ is the trivial morphism defined by $f(g) = e_H$ for all $g \in G$. □

Example 4.29. Find all the subgroups and quotient groups of D_4, the symmetry group of a square, and draw the poset diagram of its subgroups.

Solution. Any symmetry of the square induces a permutation of its vertices. Thus, as in Example 3.30, this defines a group morphism $f: D_4 \to S_4$. However, unlike the case of the symmetries of an equilateral triangle, this is not an isomorphism because $|D_4| = 8$, whereas $|S_4| = 24$. The kernel of f consists of symmetries fixing the vertices and so consists of the identity only. Therefore, by the morphism theorem, D_4 is isomorphic to the image of f in S_4. We equate an element of D_4 with its image in S_4. All the elements of D_4 are shown in Figure 4.2. The corner by the vertex 1 is blocked in, and the reverse side of the square is shaded to illustrate the effect of the symmetries. The order of each symmetry is given in Table 4.3.

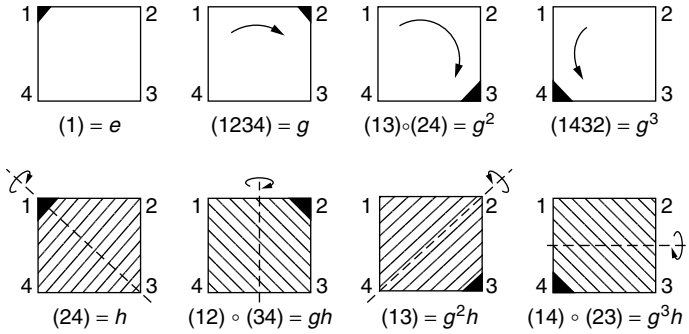


Figure 4.2. Symmetries of the square.

TABLE 4.3. Orders of the Symmetries of a Square

Elements of D_4	e	g	g^2	g^3	h	gh	g^2h	g^3h
Order of Element	1	4	2	4	2	2	2	2

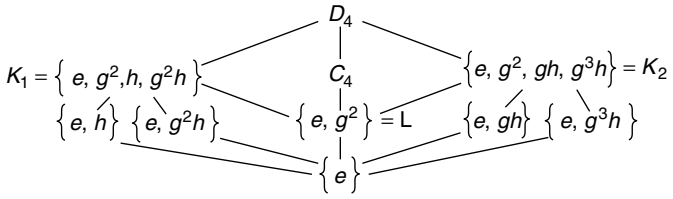


Figure 4.3. Poset diagram of subgroups of D_4.

The cyclic subgroups generated by the elements are $\{e\}$, $C_4 = \{e, g, g^2, g^3\}$, $\{e, g^2\}$, $\{e, h\}$, $\{e, gh\}$, $\{e, g^2h\}$, and $\{e, g^3h\}$.

By Lagrange's theorem, any proper subgroup must have order 2 or 4. Since any group of order 2 is cyclic, the only proper subgroups that are not cyclic are of order 4 and contain elements of order 1 and 2. There are two such subgroups, $K_1 = \{e, g^2, h, g^2h\}$ and $K_2 = \{e, g^2, gh, g^3h\}$. All the subgroups are illustrated in Figure 4.3.

To find all the quotient groups, we must determine which subgroups are normal.

The trivial group $\{e\}$ and the whole group D_4 are normal subgroups. Since C_4, K_1, and K_2 have index 2 in D_4, they are normal by Proposition 4.20, and their quotient groups are cyclic of order 2.

Subgroup H	Left Coset gH	Right Coset Hg
$\{e, h\}$	$\{g, gh\}$	$\{g, hg\} = \{g, g^3h\}$
$\{e, g^2h\}$	$\{g, g^3h\}$	$\{g, g^2hg\} = \{g, gh\}$
$\{e, gh\}$	$\{g, g^2h\}$	$\{g, ghg\} = \{g, h\}$
$\{e, g^3h\}$	$\{g, h\}$	$\{g, g^3hg\} = \{g, g^2h\}$

For each of the subgroups above, the left and right cosets containing g are different; therefore, none of the subgroups are normal.

Left Cosets of L	Right Cosets of L
$L = \{e, g^2\}$	$L = \{e, g^2\}$
$gL = \{g, g^3\}$	$Lg = \{g, g^3\}$
$hL = \{h, hg^2\} = \{h, g^2h\}$	$Lh = \{h, g^2h\}$
$ghL = \{gh, ghg^2\} = \{gh, g^3h\}$	$Lgh = \{gh, g^3h\}$

The table above shows that $L = \{e, g^2\}$ is a normal subgroup. The multiplication table for D_4/L given in Table 4.4 shows that it is isomorphic to the Klein 4-group. □

TABLE 4.4. Group D_4/L

$\cdot$	L	Lh	Lg	Lgh
L	L	Lh	Lg	Lgh
Lh	Lh	L	Lgh	Lg
Lg	Lg	Lgh	L	Lh
Lgh	Lgh	Lg	Lh	L

DIRECT PRODUCTS

Given two sets, S and T, we can form their Cartesian product, $S \times T = \{(s, t) | s \in S, t \in T\}$, whose elements are ordered pairs. For example, the product of the real line, $\mathbb{R}$, with itself is the plane, $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$. We now show how to define the product of any two groups; the underlying set of the product is the Cartesian product of the underlying sets of the original groups.

Proposition 4.30. If $(G, \circ)$ and $(H, \star)$ are two groups, then $(G \times H, \cdot)$ is a group under the operation $\cdot$ defined by

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

The group $(G \times H, \cdot)$ is called the ***direct product*** of the groups $(G, \circ)$ and $(H, \star)$.

Proof. All the group axioms follow from the axioms for $(G, \circ)$ and $(H, \star)$. The identity of $G \times H$ is (e_G, e_H), and the inverse of (g, h) is (g^{-1}, h^{-1}). □

This construction can be iterated any finite number of times to obtain the direct product of n groups.

Sometimes the direct product of two groups G and H is called the ***direct sum*** and is denoted by $G \oplus H$. (The direct sum of a finite number of groups is the same as the direct product. It is possible to define a direct sum and direct product of an infinite number of groups; these are different. An element of the direct product is obtained by taking one element from each group, while an element of the direct sum is obtained by taking one element from each group, but with only a finite number different from the identity.)

Example 4.31. Write down the table for the direct product of C_2 with itself.

Solution. Let $C_2 = \{e, g\}$, so that $C_2 \times C_2 = \{(e, e), (e, g), (g, e), (g, g)\}$. Its table is given in Table 4.5. We see that this group $C_2 \times C_2$ is isomorphic to the Klein 4-group of symmetries of a rectangle. □

Theorem 4.32. If $\gcd(m, n) = 1$, then $C_{mn} \cong C_m \times C_n$.

TABLE 4.5. Group $C_2 \times C_2$

$\cdot$	(e, e)	(e, g)	(g, e)	(g, g)
(e, e)	(e, e)	(e, g)	(g, e)	(g, g)
(e, g)	(e, g)	(e, e)	(g, g)	(g, e)
(g, e)	(g, e)	(g, g)	(e, e)	(e, g)
(g, g)	(g, g)	(g, e)	(e, g)	(e, e)

Proof. Let g, h, and k be the generators of C_{mn}, C_m, and C_n, respectively. Define

$$f: C_{mn} \to C_m \times C_n$$

by $f(g^r) = (h^r, k^r)$ for $r \in \mathbb{Z}$. This is well defined for all integers r because if $g^r = g^{r'}$, then $r - r'$ is a multiple of mn, so $r - r'$ is a multiple of m and of n. Hence $h^r = h^{r'}$ and $k^r = k^{r'}$. Now f is a group morphism because

$$\begin{aligned} f(g^r \cdot g^s) &= f(g^{r+s}) = (h^{r+s}, k^{r+s}) = (h^r \cdot h^s, k^r \cdot k^s) = (h^r, k^r) \cdot (h^s, k^s) \\ &= f(g^r) \cdot f(g^s). \end{aligned}$$

If $g^r \in \operatorname{Ker} f$, then $h^r = e$ and $k^r = e$. Therefore, r is divisible by m and n, and since $\gcd(m, n) = 1$, r is divisible by mn. Hence $\operatorname{Ker} f = \{e\}$, and the image of f is isomorphic to C_{mn}. However, $|C_{mn}| = mn$ and $|C_m \times C_n| = |C_m| \cdot |C_n| = mn$; hence $\operatorname{Im} f = C_m \times C_n$, and f is an isomorphism. □

The following is an easy consequence of this result.

Corollary 4.33. Let $n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_r^{\alpha_r}$ where $p_1, p_2, \dots, p_r$ are distinct primes. Then $C_n \cong C_{p_1^{\alpha_1}} \times C_{p_2^{\alpha_2}} \times \cdots \times C_{p_r^{\alpha_r}}$. □

If m and n are not coprime, then C_{mn} is never isomorphic to $C_m \times C_n$. For example, $C_2 \times C_2$ is not isomorphic to C_4 because the direct product contains no element of order 4. In general, the order of the element (h, k) in $H \times K$ is the least common multiple of the orders of h and k, because $(h, k)^r = (h^r, k^r) = (e, e)$ if and only if $h^r = e$ and $k^r = e$. Hence, if $\gcd(m, n) > 1$, the order of (h, k) in $C_m \times C_n$ is always less than mn.

Direct products can be used to classify all finite abelian groups. It can be shown that *any finite abelian group* is isomorphic to a *direct product of cyclic groups*. For example, see Nicholson [11] or Baumslag and Chandler [25]. The results above can be used to sort out those products of cyclic groups that are isomorphic to each other. For example, there are three nonisomorphic abelian groups with 24 elements, namely,

$$\begin{aligned} C_8 \times C_3 &\cong C_{24} \\ C_2 \times C_4 \times C_3 &\cong C_6 \times C_4 \cong C_2 \times C_{12} \\ C_2 \times C_2 \times C_2 \times C_3 &\cong C_2 \times C_2 \times C_6. \end{aligned}$$

Theorem 4.34. If $(G, \cdot)$ is a finite group for which every element $g \in G$ satisfies $g^2 = e$, then $|G| = 2^n$ for some $n \geqslant 0$, and G is isomorphic to the n-fold direct product $C_2^n = C_2 \times C_2 \times \cdots \times C_2$.

Proof. Every element in G has order 1 or 2, and the identity is the only element of order 1. Therefore, every element of G is its own inverse. The group G is abelian because for any $g, h \in G$, $gh = (gh)^{-1} = h^{-1}g^{-1} = hg$.

Choose the elements $a_1, a_2, \ldots, a_n \in G$ so that $a_i \neq e$ and a_i cannot be written as a product of powers of $a_1, \ldots, a_{i-1}$. Furthermore, choose n maximal, so that every element can be written in terms of the elements a_i. If C_2 is generated by g, we show that the function

$$f: C_2^n \to G, \text{ defined by } f(g^{r_1}, g^{r_2}, \ldots, g^{r_n}) = a_1^{r_1} a_2^{r_2} \ldots a_n^{r_n}$$

is an isomorphism. It is well defined for all integers r_i, because if $g^{r_i} = g^{q_i}$, then $a_i^{r_i} = a_i^{q_i}$. Now

$$\begin{aligned} f((g^{r_1}, \ldots, g^{r_n}) \cdot (g^{s_1}, \ldots, g^{s_n})) &= f(g^{r_1+s_1}, \ldots, g^{r_n+s_n}) = a_1^{r_1+s_1} \ldots a_n^{r_n+s_n} \\ &= a_1^{r_1} \ldots a_n^{r_n} . a_1^{s_1} \ldots a_n^{s_n} \quad \text{because } G \text{ is abelian} \\ &= f(g^{r_1}, \ldots, g^{r_n}) \cdot f(g^{s_1}, \ldots, g^{s_n}). \end{aligned}$$

Hence f is a group morphism.

Let $(g^{r_1}, \ldots, g^{r_n}) \in \operatorname{Ker} f$. Suppose that r_i is the last odd exponent, so that $r_{i+1}, r_{i+2}, \ldots, r_n$ are all even. Then $a_1^{r_1 \ldots} a_{i-1}^{r_{i-1}} a_i = e$ and

$$a_i = a_i^{-1} = a_1^{r_1 \ldots} a_{i-1}^{r_{i-1}},$$

which is a contradiction. Therefore, all the exponents are even, and f is injective. The choice of the elements a_i guarantees that f is surjective. Hence f is the required isomorphism. □

Example 4.35. Describe all the group morphisms from C_{10} to $C_2 \times C_5$. Which of these are isomorphisms? □

Solution. Since C_{10} is a cyclic group, generated by g, for example, a morphism from C_{10} is determined by the image of g. Let h and k be generators of C_2 and C_5, respectively, and consider the function $f_{r,s}: C_{10} \to C_2 \times C_5$ which maps g to the element $(h^r, k^s) \in C_2 \times C_5$. Then, if $f_{r,s}$ is a morphism, $f_{r,s}(g^n) = (h^{rn}, k^{sn})$ for $0 \leqslant n \leqslant 9$. However, this would also be true for all integers n, because if $g^n = g^m$, then $10|n - m$. Hence $2|n - m$ and $5|n - m$ and $h^{rn} = h^{rm}$ and $k^{sn} = k^{sm}$.

We now verify that $f_{r,s}$ is a morphism for any r and s. We have

$$\begin{aligned} f_{r,s}(g^a g^b) = f_{r,s}(g^{a+b}) &= (h^{(a+b)r}, k^{(a+b)s}) = (h^{ar}, k^{as})(h^{br}, k^{bs}) \\ &= f_{r,s}(g^a) f_{r,s}(g^b). \end{aligned}$$

Therefore, there are ten morphisms, $f_{r,s}$, from C_{10} to $C_2 \times C_5$ corresponding to the ten elements (h^r, k^s) of $C_2 \times C_5$.

Now

$$\text{Ker} f_{r,s} = \{g^n | (h^{rn}, k^{sn}) = (e, e)\} = \{g^n | rn \equiv 0 \bmod 2 \text{ and } sn \equiv 0 \bmod 5\}.$$

Hence $\text{Ker} f_{r,s} = \{e\}$ if $(r, s) = (1, 1), (1, 2), (1, 3)$, or $(1, 4)$, while $\text{Ker} f_{0,0} = C_{10}$, $\text{Ker} f_{1,0} = \{e, g^2, g^4, g^6, g^8\}$, and $\text{Ker} f_{0,s} = \{e, g^5\}$, if $s = 1, 2, 3$, or 4. If $\text{Ker} f_{r,s}$ contains more than one element, $f_{r,s}$ is not an injection and cannot be an isomorphism. By the morphism theorem,

$$|C_{10}|/|\text{Ker} f_{r,s}| = |\text{Im} f_{r,s}|,$$

and if $\text{Ker} f_{r,s} = \{e\}$, then $|\text{Im } f_{r,s}| = 10$, so $f_{r,s}$ is surjective also. Therefore, the isomorphisms are $f_{1,1}$, $f_{1,2}$, $f_{1,3}$, and $f_{1,4}$. □

GROUPS OF LOW ORDER

We find all possible isomorphism classes of groups with eight or fewer elements.

Lemma 4.36. Suppose that a and b are elements of coprime orders r and s, respectively, in an abelian group. Then ab has order rs.

Proof. Let A and B denote the subgroups generated by a and b, respectively. Since $ab = ba$, we have $(ab)^{rs} = a^{rs}b^{rs} = (a^r)^s(b^s)^r = e^s e^r = e$. Suppose that $(ab)^k = e$; we must show that rs divides k. Observe that $a^k = b^{-k} \in A \cap B$. Since $A \cap B$ is a subgroup of both A and B, its order divides $|A| = r$ and $|B| = s$ by Lagrange's theorem. Since r and s are coprime, this implies that $|A \cap B| = 1$. It follows that $a^k = e$ and $b^{-k} = e$, so r divides k and s divides k. Hence rs divides k by Theorem 11, Appendix 2 (again because r and s are coprime), as required. □

With this we can describe the groups of order eight or less.

Order 1. Every trivial group is isomorphic to $\{e\}$.

Order 2. By Corollary 4.11, every group of order 2 is cyclic.

Order 3. By Corollary 4.11, every group of order 3 is cyclic.

Order 4. Each element has order 1, 2, or 4.

Case (i). If there is an element of order 4, the group is cyclic.

Case (ii). If not, every element has order 1 or 2 and, by Theorem 4.34, the group is isomorphic to $C_2 \times C_2$.

Order 5. By Corollary 4.11, every group of order 5 is cyclic.

Order 6. Each element has order 1, 2, 3, or 6.

Case (i). If there is an element of order 6, the group is cyclic.

Case (ii). If not, the elements have orders 1, 2, or 3. By Theorem 4.34, all the elements in a group of order 6 cannot have orders 1 and 2. Hence there is an element, say a, of order 3. The subgroup $H = \{e, a, a^2\}$ has index 2, and if $b \notin H$, the underlying set of the group is then $H \cup Hb = \{e, a, a^2, b, ab, a^2b\}$. By Proposition 4.20, H is normal, and the quotient group of H is cyclic of order 2. Hence

$$b^r \in Hb^r = (Hb)^r = \begin{cases} H & \text{if } r \text{ is even} \\ Hb & \text{if } r \text{ is odd} \end{cases}.$$

Therefore, b has even order. It cannot be 6, so it must be 2. As H is normal, $bab^{-1} \in H$. We cannot have $bab^{-1} = e$, because $a \neq e$. If $bab^{-1} = a$, then $ba = ab$, and we can prove that the entire group is abelian. This cannot happen because by Lemma 4.36, ab would have order 6. Therefore, $bab^{-1} = a^2$, and the group is generated by a and b with relations $a^3 = b^2 = e$ and $ba = a^2b$. This group is isomorphic to D_3 and S_3.

Order 7. Every group of order 7 is cyclic.

Order 8. Each element has order 1, 2, 4, or 8.

Case (i). If there is an element of order 8, the group is cyclic.

Case (ii). If all elements have order 1 or 2, the group is isomorphic to $C_2 \times C_2 \times C_2$ by Theorem 4.34.

Case (iii). Otherwise, there is an element of order 4, say a. The subgroup $H = \{e, a, a^2, a^3\}$ is of index 2 and therefore normal. If $b \notin H$, the underlying set of the group is $H \cup Hb = \{e, a, a^2, a^3, b, ab, a^2b, a^3b\}$. Now $b^2 \in H$, but b^2 cannot have order 4; otherwise, b would have order 8. Therefore, $b^2 = e$ or a^2. As H is normal, $bab^{-1} \in H$ and has the same order as a because $(bab^{-1})^k = ba^kb^{-1}$.

Case (iiia). If $bab^{-1} = a$, then $ba = ab$, and the whole group can be proved to be abelian. If $b^2 = e$, each element can be written uniquely in the form a^rb^s, where $0 \leqslant r \leqslant 3$ and $0 \leqslant s \leqslant 1$. Hence the group is isomorphic to $C_4 \times C_2$ by mapping a^rb^s to (a^r, b^s). If $b^2 = a^2$, let $c = ab$, so that $c^2 = a^2b^2 = a^4 = e$. Each element of the group can now be written uniquely in the form a^rc^s, where $0 \leqslant r \leqslant 3, 0 \leqslant s \leqslant 1$, and the group is still isomorphic to $C_4 \times C_2$.

Case (iiib). If $bab^{-1} = a^3$ and $b^2 = e$, the group is generated by a and b with the relations $a^4 = b^2 = e, ba = a^3b$. This is isomorphic to the dihedral group D_4.

Case (iiic). If $bab^{-1} = a^3$ and $b^2 = a^2$, then the group is isomorphic to the quaternion group Q, described in Exercise 3.47. The isomorphism maps a^rb^s to i^rj^s.

Any group with eight or fewer elements is isomorphic to exactly one group in Table 4.6.

TABLE 4.6. Groups of Low Order

Order	1	2	3	4	5	6	7	8
Abelian groups	$\{e\}$	C_2	C_3	C_4	C_5	C_6	C_7	C_8
				$C_2 \times C_2$				$C_4 \times C_2$
								$C_2 \times C_2 \times C_2$
Non-abelian groups						S_3		D_4
								Q

ACTION OF A GROUP ON A SET

The concept of a group acting on a set X is a slight generalization of the group of symmetries of X. It is equivalent to considering a subgroup of $S(X)$. This concept is useful for determining the order of the symmetry groups of solids in three dimensions, and it is indispensable in Chapter 6, when we look at the Pólya–Burnside method of enumerating sets with symmetries.

The group $(G, \cdot)$ **acts on the set** X if there is a function

$$\psi: G \times X \to X$$

such that when we write $g(x)$ for $\psi(g, x)$, we have:

(i) $(g_1 g_2)(x) = g_1(g_2(x))$ for all $g_1, g_2 \in G, x \in X$.
(ii) $e(x) = x$ if e is the identity of G and $x \in X$.

Proposition 4.37. If g is an element of a group G acting on the set X, then the function $g: X \to X$, which maps x to $g(x)$, is a bijection. This defines a morphism

$$\chi: G \to S(X)$$

from G to the group of symmetries of X.

Proof. The function $g: X \to X$ is injective because if $g(x) = g(y)$, then $g^{-1}g(x) = g^{-1}g(y)$, and $e(x) = e(y)$ or $x = y$. It is surjective because if $z \in X$,

$$g(g^{-1}(z)) = gg^{-1}(z) = e(z) = z.$$

Hence g is bijective, and g can be considered as an element of $S(X)$, the group of symmetries of X.

The function $\chi: G \to S(X)$, which takes the element $g \in G$ to the bijection $g: X \to X$, is a group morphism because $\chi(g_1 g_2)$ is the function from X to X defined by $\chi(g_1 g_2)(x) = (g_1 g_2)(x) = (g_1(g_2(x)) = \chi(g_1) \circ \chi(g_2)(x)$; thus $\chi(g_1 g_2) = \chi(g_1) \circ \chi(g_2)$. □

If $\mathrm{Ker}\chi = \{e\}$, then χ is injective, and the group G is said to **act faithfully** on the set X. G acts faithfully on X if the only element of G, which fixes every

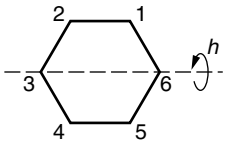


Figure 4.4. C_2 acting on a hexagon.

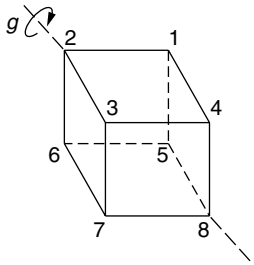


Figure 4.5. C_3 acting on a cube.

element of X, is the identity $e \in G$. In this case, we identify G with $\mathrm{Im}\chi$ and regard G as a subgroup of $S(X)$.

For example, consider the cyclic group of order 2, $C_2 = \{e, h\}$, acting on the regular hexagon in Figure 4.4, where h reflects the hexagon about the line joining vertex 3 to vertex 6. Then C_2 acts faithfully and can be identified with the subgroup $\{(1), (15) \circ (24)\}$ of D_6.

The cyclic group $C_3 = \{e, g, g^2\}$ acts faithfully on the cube in Figure 4.5, where g rotates the cube through one-third of a revolution about a line joining two opposite vertices. This group action can be considered as the subgroup $\{(1), (163) \circ (457), (136) \circ (475)\}$ of the symmetry group of the cube.

Proposition 4.38. If G acts on a set X and $x \in X$, then

$$\mathrm{Stab}\ x = \{g \in G | g(x) = x\}$$

is a subgroup of G, called the **stabilizer** of x. It is the set of elements of G that fix x.

Proof. Stab x is a subgroup because

(i) If $g_1, g_2 \in \mathrm{Stab}\ x$, then $(g_1 g_2)(x) = g_1(g_2(x)) = g_1(x) = x$, so $g_1 g_2 \in \mathrm{Stab}\ x$.
(ii) If $g \in \mathrm{Stab}\ x$, then $g^{-1}(x) = x$, so $g^{-1} \in \mathrm{Stab}\ x$. □

The set of all images of an element $x \in X$ under the action of a group G is called the ***orbit*** of x under G and is denoted by

$$\mathrm{Orb}\ x = \{g(x) | g \in G\}.$$

The orbit of x is the equivalence class of x under the equivalence relation on X in which x is equivalent to y if and only if $y = g(x)$ for some $g \in G$. If π is a permutation in S_n, the subgroup generated by π acts on the set $\{1, 2, \ldots, n\}$, and this definition of orbit agrees with our previous one.

A graphic illustration of orbits can be obtained by looking at the group of matrices

$$\mathrm{SO}(2) = \left\{ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \middle| \theta \in \mathbb{R} \right\}$$

under matrix multiplication. This group is called the **special orthogonal group** and is isomorphic to the circle group W. SO(2) acts on $\mathbb{R}^2$ as follows. The matrix $M \in \mathrm{SO}(2)$ takes the vector $\mathbf{x} \in \mathbb{R}^2$ to the vector $M\mathbf{x}$. The orbit of any element $\mathbf{x} \in \mathbb{R}^2$ is the circle through $\mathbf{x}$ with center at the origin. Since the origin is the only fixed point for any of the nonidentity transformations, the stabilizer of the origin is the whole group, whereas the stabilizer of any other element is the subgroup consisting of the identity matrix only.

The orbits of the cyclic group C_2 acting on the hexagon in Figure 4.4 are $\{1, 5\}$, $\{2, 4\}$, $\{3\}$, and $\{6\}$.

There is an important connection between the number of elements in the orbit of a point x and the stabilizer of that point.

Lemma 4.39. If G acts on X, then for each $x \in X$

$$|G\colon \mathrm{Stab}\ x| = |\mathrm{Orb}\ x|.$$

Proof. Let $H = \mathrm{Stab}\ x$ and define the function

$$\xi\colon G/H \to \mathrm{Orb}\ x$$

by $\xi(Hg) = g^{-1}(x)$. This is well defined on cosets because if $Hg = Hk$, then $k = hg$ for some $h \in H$, so $k^{-1}(x) = (hg)^{-1}(x) = g^{-1}h^{-1}(x) = g^{-1}(x)$, since $h^{-1} \in H = \mathrm{Stab}\ x$.

The function ξ is surjective by the definition of the orbit of x. It is also injective, because $\xi(Hg_1) = \xi(Hg_2)$ implies that $g_1^{-1}(x) = g_2^{-1}(x)$, so $g_2g_1^{-1}(x) = x$ and $g_2g_1^{-1} \in \mathrm{Stab}\ x = H$. Therefore, ξ is a bijection, and the result follows. □

Note that ξ is *not* a morphism. $G/\mathrm{Stab}\ x$ is just a set of cosets because Stab x is not necessarily normal. Furthermore, we have placed no group structure on Orb x.

Theorem 4.40. If the finite group G acts on a set X, then for each $x \in X$,

$$|G| = |\mathrm{Stab}\ x||\mathrm{Orb}\ x|.$$

Proof. This follows from Lemma 4.39 and Corollary 4.8. □

Example 4.41. Find the number of proper rotations of a cube.

Solution. Let G be the group of proper rotations of a cube; that is, rotations that can be carried out in three dimensions. The stabilizer of the vertex 1 in Figure 4.6 is $\mathrm{Stab}\ 1 = \{(1), (245) \circ (386), (254) \circ (368)\}$. The orbit of 1 is the set of all the vertices, because there is an element of G that will take 1 to any other vertex. Therefore, by Theorem 4.40,

$$|G| = |\mathrm{Stab}\ 1|\ |\mathrm{Orb}\ 1| = 3 \cdot 8 = 24.$$ □

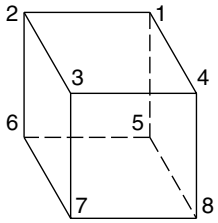


Figure 4.6. Cube.

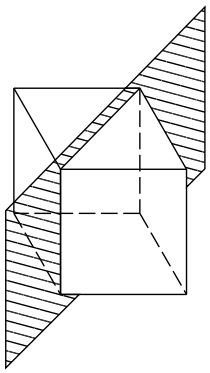


Figure 4.7. Reflection in a plane.

The full symmetry group of the cube would include improper rotations such as the reflection in the plane as shown in Figure 4.7. This induces the permutation $(24) \circ (68)$ on the vertices, and it cannot be obtained by physically rotating the cube in three dimensions. Under this group

$$\text{Stab } 1 = \{(1), (245) \circ (368), (254) \circ (386), (24) \circ (68), (25) \circ (38), (45) \circ (36)\},$$

so the order of the full symmetry group of the cube is

$$|\text{Stab } 1||\text{Orb } 1| = 6.8 = 48.$$

Therefore, there are 24 proper and 24 improper rotations of the cube.

The article by Shapiro [32] contains many applications, mainly to group theory, of the actions of a group on a set.

We conclude by mentioning the action of the symmetric group S_n on the set of polynomials in n variables $x_1, x_2, \ldots, x_n$. A permutation $\sigma \in S_n$ acts on a polynomial $f = f(x_1, x_2, \ldots, x_n)$ by permuting the variables:

$$\sigma(f) = f(x_{\sigma 1}, x_{\sigma 2}, \ldots, x_{\sigma n}).$$

This was the action we used in Chapter 3 to define the parity of a permutation and prove the parity theorem. It has historical interest as well. It is the context in which Lagrange proved Lagrange's theorem—in essence, what he actually did is prove Theorem 4.40 for this action. Moreover, this is the group action that launched Galois theory, about which we say more in Chapter 11.

EXERCISES

In Exercises 4.1 to 4.4, which of the relations are equivalence relations? Describe the equivalence classes of those relations which are equivalence relations.

4.1. The relation $\sim$ on $\mathbb{P} \times \mathbb{P}$ defined by $(a, b) \sim (c, d)$ if and only if $a + d = b + c$.

4.2. The relation T on the set of continuous functions from $\mathbb{R}$ to $\mathbb{R}$, where fTg if and only if $f(3) = g(3)$.

4.3. The inclusion relation on the power set $\mathcal{P}(X)$.

4.4. The relation C on a group G, where aCb if and only if $ab = ba$.

4.5. Find the left and right cosets of $H = \{(1), (12), (34), (12) \circ (34)\}$ in S_4.

4.6. Let H be the subgroup of A_4 that fixes 1. Find the left and right cosets of H in A_4. Is H normal? Describe the left cosets in terms of their effect on the element 1. Can you find a similar description for the right cosets?

In Exercises 4.7 to 4.12, verify that each of the functions is well defined. Determine which are group morphisms, and find the kernels and images of all the morphisms. The element of $\mathbb{Z}_n$ containing x is denoted by $[x]_n$.

4.7. $f: \mathbb{Z}_{12} \to \mathbb{Z}_{12}$, where $f([x]_{12}) = [x + 1]_{12}$.

4.8. $f: C_{12} \to C_{12}$, where $f(g) = g^3$.

4.9. $f: \mathbb{Z} \to \mathbb{Z}_2 \times \mathbb{Z}_4$, where $f(x) = ([x]_2, [x]_4)$.

4.10. $f: \mathbb{Z}_8 \to \mathbb{Z}_2$, where, $f([x]_8) = [x]_2$.

4.11. $f: C_2 \times C_3 \to C_3$, where $f(h^r, k^s) = (12)^r \circ (123)^s$.

4.12. $f: S_n \to S_{n+1}$, where $f(\pi)$ is the permutation on $\{1, 2, \ldots, n + 1\}$ defined by $f(\pi)(i) = \pi(i)$ if $i \leqslant n$, and $f(\pi)(n + 1) = n + 1$.

4.13. If H is a subgroup of an abelian group G, prove that the quotient group G/H is abelian.

4.14. If H is a subgroup of G, show that $g^{-1}Hg = \{g^{-1}hg | h \in H\}$ is a subgroup for each $g \in G$.

4.15. Prove that the subgroup H is normal in G if and only if $g^{-1}Hg = H$ for all $g \in G$.

4.16. If H is the only subgroup of a given order in a group G, prove that H is normal in G.

4.17. Let H be any subgroup of a group G. Prove that there is a one-to-one correspondence between the set of left cosets of H in G and the set of right cosets of H in G.

4.18. Is the cyclic subgroup $\{(1), (123), (132)\}$ normal in S_4?

4.19. Is the cyclic subgroup $\{(1), (123), (132)\}$ normal in A_4?

4.20. Is $\{(1), (1234), (13) \circ (24), (1432), (13), (24), (14) \circ (23), (12) \circ (34)\}$ normal in S_4?

4.21. Find all the group morphisms from C_3 to C_4.

4.22. Find all the group morphisms from $\mathbb{Z}$ to $\mathbb{Z}_4$.

4.23. Find all the group morphisms from C_6 to C_6.

4.24. Find all the group morphisms from $\mathbb{Z}$ to D_4.

In Exercises 4.25 to 4.29, which of the pairs of groups are isomorphic? Give reasons.

4.25. C_{60} and $C_{10} \times C_6$.

4.26. $(\mathcal{P}\{a, b, c\}, \Delta)$ and $C_2 \times C_2 \times C_2$.

4.27. D_n and $C_n \times C_2$.

4.28. D_6 and A_4.

4.29. $\mathbb{Z}_4 \times \mathbb{Z}_2$ and $(\{\pm 1, \pm i, \pm(1+i)/\sqrt{2}, \pm(1-i)/\sqrt{2}\}, \cdot)$.

4.30. If $G \times H$ is cyclic, prove that G and H are cyclic.

4.31. If π is an r-cycle in S_n, prove that $\rho^{-1} \circ \pi \circ \rho$ is also an r-cycle for each $\rho \in S_n$.

4.32. Find four different subgroups of S_4 that are isomorphic to S_3.

4.33. Find all the isomorphism classes of groups of order 10.

4.34. Find all the ten subgroups of A_4 and draw the poset diagram under inclusion. Which of the subgroups are normal?

4.35. For any groups G and H, prove that $(G \times H)/G' \cong H$ and $(G \times H)/H' \cong G$, where $G' = \{(g, e) \in G \times H | g \in G\}$ and $H' = \{(e, h) \in G \times H | h \in H\}$.

4.36. Show that $\mathbb{Q}/\mathbb{Z}$ is an infinite group but that every element has finite order.

4.37. If G is a subgroup of S_n and G contains an odd permutation, prove that G contains a normal subgroup of index 2.

4.38. In any group $(G, \cdot)$ the element $a^{-1}b^{-1}ab$ is called the **commutator** of a and b. Let G' be the subset of G consisting of all finite products of commutators. Show that G' is a normal subgroup of G. This is called the **commutator subgroup**. Also prove that G/G' is abelian.

4.39. Let $\mathbb{C}^*$ be the group of nonzero complex numbers under multiplication and let W be the multiplicative group of complex numbers of unit modulus. Describe $\mathbb{C}^*/W$.

4.40. Show that $K = \{(1), (12) \circ (34), (13) \circ (24), (14) \circ (23)\}$ is a subgroup of S_4 isomorphic to the Klein 4-group. Prove that K is normal and that $S_4/K \cong S_3$.

4.41. If K is the group given in Exercise 4.40, prove that K is normal in A_4 and that $A_4/K \cong C_3$. This shows that A_4 is not a simple group.

4.42. The *cross-ratio* of the four distinct real numbers x_1, x_2, x_3, x_4 in that order is the ratio $\lambda = (x_2 - x_4)(x_3 - x_1)/(x_2 - x_1)(x_3 - x_4)$. Find the subgroup K of S_4, of all those permutations of the four numbers that preserve the value of the cross-ratio. Show that if λ is the cross-ratio of four numbers taken in a certain order, the cross-ratio of these numbers in any other order must belong to the set

$$\left\{\lambda, 1-\lambda, \frac{1}{\lambda}, 1-\frac{1}{\lambda}, \frac{1}{1-\lambda}, \frac{\lambda}{\lambda-1}\right\}.$$

Furthermore, show that all permutations in the same coset of K in S_4 give rise to the same cross-ratio. In other words, prove that the quotient group S_4/K is isomorphic to the group of functions given in Exercise 3.42. The cross-ratio is very useful in projective geometry because it is preserved under projective transformations.

4.43. (**Second Isomorphism Theorem**) Let N be a normal subgroup of G, and let H be any subgroup of G. Show that $HN = \{hn | h \in H, n \in N\}$ is a subgroup of G and that $H \cap N$ is a normal subgroup of H. Also prove that

$$H/(H \cap N) \cong HN/N.$$

4.44. (**Third isomorphism theorem**) Let M and N be normal subgroups of G, and N be a normal subgroup of M. Show that $\phi: G/N \to G/M$ is a well-defined morphism if $\phi(Ng) = Mg$, and prove that

$$(G/N)/(M/N) \cong G/M.$$

4.45. If a finite group contains no nontrivial subgroups, prove that it is either trivial or cyclic of prime order.

4.46. If d is a divisor of the order of a finite cyclic group G, prove that G contains a subgroup of order d.

4.47. If G is a finite abelian group and p is a prime such that $g^p = e$, for all $g \in G$, prove that G is isomorphic to $\mathbb{Z}_p^n$, for some integer n.

4.48. What is the symmetry group of a rectangular box with sides of length 2, 3, and 4 cm?

4.49. Let

$$G_p = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}_p) | ad - bc = 1 \text{ in } \mathbb{Z}_p \right\}.$$

If p is prime, show that $(G_p, \cdot)$ is a group of order $p(p^2 - 1)$, and find a group isomorphic to G_2.

4.50. Show that $(\mathbb{R}^*, \cdot)$ acts on $\mathbb{R}^{n+1}$ by scalar multiplication. What are the orbits under this action? The set of orbits, excluding the origin, form the n-dimensional real *projective space*.

4.51. Let G be a group of order n, and let $\gcd(n, m) = 1$. Show that every element h in G has an mth root; that is, $h = g^m$ for some $g \in G$.

4.52. Let G' denote the commutator subgroup of a group G (see Exercise 4.38). If K is a subgroup of G, show that $G' \subseteq H$ if and only if K is normal in G and G/K is abelian.

4.53. Call a group G **metabelian** if it has a normal subgroup K such that both K and G/K are abelian.

(a) Show that every subgroup and factor group of a metabelian group is metabelian. (Exercises 4.43 and 4.44 are useful.)

(b) Show that G is metabelian if and only if the commutator group G' is abelian (see Exercise 4.38).

4.54. Recall (Exercise 3.76) that the center $Z(G)$ of a group G is defined by $Z(G) = \{z \in G | zg = gz \text{ for all } g \in G\}$. Let $K \subseteq Z(G)$ be a subgroup.
(a) Show that K is normal in G.
(b) If G/K is cyclic, show that G is abelian.

For Exercises 4.55 to 4.61, let $\mathbb{Z}_m{}^* = \{[x] \in \mathbb{Z}_m | gcd(x, m) = 1\}$. *The number of elements in this set is denoted by* $\phi(m)$ *and is called the* **Euler ϕ-function**. *For example,* $\phi(4) = 2$, $\phi(6) = 2$, *and* $\phi(8) = 4$.

4.55. Show that $\phi(p^r) = p^r - p^{r-1}$ if p is a prime.

4.56. Show that $\phi(mn) = \phi(m)\phi(n)$ if $\gcd(m, n) = 1$.

4.57. Prove that $(\mathbb{Z}_m{}^*, \cdot)$ is an abelian group.

4.58. Write out the multiplication table for $(\mathbb{Z}_8{}^*, \cdot)$.

4.59. Prove that $(\mathbb{Z}_6{}^*, \cdot)$ and $(\mathbb{Z}_{17}{}^*, \cdot)$ are cyclic and find generators.

4.60. Find groups in Table 4.6 that are isomorphic to $(\mathbb{Z}_8{}^*, \cdot)$, $(\mathbb{Z}_9{}^*, \cdot)$, $(\mathbb{Z}_{10}{}^*, \cdot)$, and $(\mathbb{Z}_{15}{}^*, \cdot)$ and describe the isomorphisms.

4.61. Prove that if $\gcd(a, m) = 1$, then $a^{\phi(m)} \equiv 1 \bmod m$. [This result was known to Leonhard Euler (1707–1783).]

4.62. Prove that if p is a prime, then for any integer a, $a^p \equiv a \bmod p$. [This result was known to Pierre de Fermat (1601–1665).]

4.63. If G is a group of order 35 acting on a set with 13 elements, show that G must have a fixed point, that is, a point $x \in S$ such that $g(x) = x$ for all $g \in G$.

4.64. If G is a group of order p^r acting on a set with m elements, show that G has a fixed point if p does not divide m.

5

SYMMETRY GROUPS IN THREE DIMENSIONS

In this chapter we determine the symmetry groups that can be realized in two- and three-dimensional space. We rely heavily on geometric intuition, not only to simplify arguments but also to give geometric flavor to the group theory. Because we live in a three-dimensional world, these symmetry groups play a crucial role in the application of modern algebra to physics and chemistry.

We first show how the group of isometries of $\mathbb{R}^n$ can be broken down into translations and orthogonal transformations fixing the origin. Since the orthogonal transformations can be represented as a group of matrices, we look at the properties of matrix groups. We then use these matrix groups to determine all the finite rotation groups in two and three dimensions, and we find polyhedra that realize these symmetry groups.

TRANSLATIONS AND THE EUCLIDEAN GROUP

Euclidean geometry in n dimensions is concerned with those properties that are preserved under isometries (rigid motions) of euclidean n-space, that is, bijections $\alpha: \mathbb{R}^n \to \mathbb{R}^n$ that preserve distance. The group of all isometries of $\mathbb{R}^n$ is called the **euclidean group** in n dimensions and is denoted $E(n)$. Given $\mathbf{w} \in \mathbb{R}^n$, the map $\mathbb{R}^n \to \mathbb{R}^n$ with $\mathbf{v} \mapsto \mathbf{v} + \mathbf{w}$ is called **translation** by $\mathbf{w}$, and we begin by showing that the group $T(n)$ of all translations is a normal subgroup of $E(n)$, and that the factor group is isomorphic to the group of all orthogonal $n \times n$ matrices (that is, matrices A such that $A^{-1} = A^T$, the **transpose** of A—reflection of A in its main diagonal).

Recall that a function $\lambda: \mathbb{R}^n \to \mathbb{R}^n$ is called a **linear transformation** if $\lambda(a\mathbf{v} + b\mathbf{w}) = a\lambda(\mathbf{v}) + b\lambda(\mathbf{w})$ for all $a, b \in \mathbb{R}$ and all (column) vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. Let $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ denote the **standard basis** of $\mathbb{R}^n$, that is, the columns of the

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

$n \times n$ identity matrix. Then the action of λ is matrix multiplication $\lambda(\mathbf{v}) = A\mathbf{v}$ for all $\mathbf{v}$ in $\mathbb{R}^n$, where the matrix A is given in terms of its columns by $A = [\lambda(\mathbf{e}_1)\lambda(\mathbf{e}_2)\cdots\lambda(\mathbf{e}_n)]$ and is called the **standard matrix** of α. Moreover, the correspondence $\lambda \leftrightarrow A$ is a bijection that preserves addition, multiplication, and the identity. So we may (and sometimes shall) identify λ with the matrix A.

If $\mathbf{v}$ and $\mathbf{w}$ are vectors in $\mathbb{R}^n$, let $\mathbf{v}\bullet\mathbf{w} = \mathbf{v}^T\mathbf{w}$ denote their **inner product**. Then $\|\mathbf{v}\| = \sqrt{\mathbf{v}\bullet\mathbf{v}}$ is the **length** of $\mathbf{v}$, and $\|\mathbf{v} - \mathbf{w}\|$ is the **distance between** $\mathbf{v}$ and $\mathbf{w}$. Thus a function $\alpha: \mathbb{R}^n \to \mathbb{R}^n$ is an isometry if

$$\|\alpha(\mathbf{v}) - \alpha(\mathbf{w})\| = \|\mathbf{v} - \mathbf{w}\| \quad \text{for all } \mathbf{v}, \mathbf{w} \in \mathbb{R}^n. \tag{*}$$

Since $\|\mathbf{v} - \mathbf{w}\|^2 = \|\mathbf{v}\|^2 + 2(\mathbf{v}\bullet\mathbf{w}) + \|\mathbf{w}\|^2$ for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, it follows from (*) that every isometry α preserves inner products in the sense that

$$\alpha(\mathbf{v})\bullet\alpha(\mathbf{w}) = \mathbf{v}\bullet\mathbf{w} \quad \text{for all } \mathbf{v}, \mathbf{w} \in \mathbb{R}^n. \tag{**}$$

Lemma 5.1. If $\alpha: \mathbb{R}^n \to \mathbb{R}^n$ is an isometry such that $\alpha(0) = 0$, then α is linear.

Proof. It follows from (**) that $\{\alpha(\mathbf{e}_1), \alpha(\mathbf{e}_2), \ldots, \alpha(\mathbf{e}_n)\}$ is an orthonormal basis of $\mathbb{R}^n$. If $a \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^n$, then (**) implies that

$$[\alpha(a\mathbf{v}) - a\alpha(\mathbf{v})]\bullet\alpha(\mathbf{e}_i) = (a\mathbf{v})\bullet\mathbf{e}_i - a(\mathbf{v}\bullet\mathbf{e}_i) = 0 \quad \text{for each } i.$$

Hence $\alpha(a\mathbf{v}) = a\alpha(\mathbf{v})$, and $\alpha(\mathbf{v} + \mathbf{w}) = \alpha(\mathbf{v}) + \alpha(\mathbf{w})$ follows in the same way for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. □

Hence the isometries of $\mathbb{R}^n$ that fix the origin are precisely the linear isometries.

An $n \times n$ matrix A is called **orthogonal** if it is invertible and $A^{-1} = A^T$, equivalently if the columns of A are an orthonormal basis of $\mathbb{R}^n$. These matrices form a subgroup of the group of all invertible matrices, called the **orthogonal group** and denoted $O(n)$.

Proposition 5.2. Let $\lambda: \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation with standard matrix A.

(i) λ is an isometry if and only if A is an orthogonal matrix.

(ii) The group of linear isometries of $\mathbb{R}^n$ is isomorphic to $O(n)$.

Proof. If A is orthogonal, then for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$,

$$\|\lambda(\mathbf{v}) - \lambda(\mathbf{w})\|^2 = A(\mathbf{v} - \mathbf{w})\bullet A(\mathbf{v} - \mathbf{w}) = (\mathbf{v} - \mathbf{w})^T A^T A(\mathbf{v} - \mathbf{w}) = \|\mathbf{v} - \mathbf{w}\|^2$$

and it follows that λ is an isometry. Conversely, if λ is an isometry and $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is the standard basis of $\mathbb{R}^n$, then (**) gives

$$\mathbf{e}_i\bullet\mathbf{e}_j = \lambda(\mathbf{e}_i)\bullet\lambda(\mathbf{e}_j) = A\mathbf{e}_i\bullet A\mathbf{e}_j = \mathbf{e}_i^T(A^T A)\mathbf{e}_j = \text{the } (i, j)\text{-entry of } A$$

for all i and j. It follows that $A^T A = I$, so A is orthogonal, proving (i). But then the correspondence $\lambda \leftrightarrow A$ between the linear transformation λ and its standard matrix A induces a group isomorphism between the (linear) isometries fixing the origin and the orthogonal matrices. This proves (ii). □

Given a vector $\mathbf{w} \in \mathbb{R}^n$, define $\tau_{\mathbf{w}}: \mathbb{R}^n \to \mathbb{R}^n$ by $\tau_{\mathbf{w}}(\mathbf{v}) = \mathbf{v} - \mathbf{w}$ for all $\mathbf{v} \in \mathbb{R}^n$. Thus $\tau_{\mathbf{w}}$ is the unique translation that carries $\mathbf{w}$ to $\mathbf{0}$. Because

$$\tau_{\mathbf{w}} \circ \tau_{\mathbf{w}'} = \tau_{\mathbf{w}+\mathbf{w}'} \quad \text{for all } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^n,$$

the correspondence $\mathbf{w} \leftrightarrow \tau_{\mathbf{w}}$ is a group isomorphism $(\mathbb{R}^n, +) \cong T(n)$. In particular, $T(n)$ is an abelian group.

Theorem 5.3. For each $n \geqslant 1$, $T(n)$ is an abelian normal subgroup of $E(n)$ and $E(n)/T(n) \cong O(n)$. In fact, the map $E(n) \to O(n)$ given by

$$\alpha \mapsto \text{the standard matrix of } \tau_{\alpha(\mathbf{0})} \circ \alpha$$

is a surjective group morphism $E(n) \to O(n)$ with kernel $T(n)$.

Proof. Write $G(n)$ for the group of all linear isometries of $\mathbb{R}^n$. Observe that if $\alpha \in E(n)$, then $\tau_{\alpha(\mathbf{0})} \circ \alpha$ is linear (it is an isometry that fixes $\mathbf{0}$), so we have a map

$$\phi: E(n) \to G(n) \quad \text{given by} \quad \phi(\alpha) = \tau_{\alpha(\mathbf{0})} \circ \alpha \quad \text{for all } \alpha \in E(n).$$

By Proposition 5.2 it suffices to show that ϕ is a surjective group morphism with kernel $T(n)$. To see that ϕ is a group morphism, observe first that

$$\alpha \circ \tau_{\mathbf{w}} = \tau_{\alpha(\mathbf{w})} \circ \alpha \quad \text{for all } \mathbf{w} \in \mathbb{R}^n. \qquad (***)$$

[Indeed, $(\alpha \circ \tau_{\mathbf{w}})(\mathbf{v}) = \alpha(\mathbf{v} - \mathbf{w}) = \alpha(\mathbf{v}) - \alpha(\mathbf{w}) = (\tau_{\alpha(\mathbf{w})} \circ \alpha)(\mathbf{v})$ for all $\mathbf{v}$.] Hence, given α and β in $E(n)$, we have

$$\phi(\alpha) \circ \phi(\beta) = \tau_{\alpha(\mathbf{0})} \circ \alpha \circ \tau_{\beta(\mathbf{0})} \circ \beta = \tau_{\alpha(\mathbf{0})} \circ (\alpha \circ \tau_{\beta(\mathbf{0})} \circ \alpha^{-1}) \circ (\alpha \circ \beta),$$

so it suffices to show that $\tau_{\alpha(\mathbf{0})} \circ (\alpha \circ \tau_{\beta(\mathbf{0})} \circ \alpha^{-1}) = \tau_{(\alpha \circ \beta)(\mathbf{0})}$. But this follows because $\alpha \circ \tau_{\beta(\mathbf{0})} \circ \alpha^{-1}$ is a translation by (***), so $\tau_{\alpha(\mathbf{0})} \circ (\alpha \circ \tau_{\beta(\mathbf{0})} \circ \alpha^{-1})$ is the unique translation that carries $(\alpha \circ \beta)(\mathbf{0})$ to 0. Hence ϕ is a group morphism. Moreover, ϕ is surjective because $\phi(\lambda) = \lambda$ for every $\lambda \in G(n)$, and $\mathrm{Ker}(\phi) = T(n)$ because $\phi(\alpha) = 1_{\mathbb{R}^n}$ if and only if $\alpha(\mathbf{v}) = \mathbf{v} + \alpha(\mathbf{0})$ for all $v \in \mathbb{R}^n$, that is, if and only if $\alpha = \tau_{-\alpha(\mathbf{0})}$. □

If $\alpha \in E(n)$ and $\alpha(\mathbf{0}) = \mathbf{w}$, the proof of Theorem 5.3 shows that $\tau_{\mathbf{w}} \circ \alpha \in G(n)$. Hence every isometry α of $\mathbb{R}^n$ is the composition of a linear isometry $\tau_{\mathbf{w}}^{-1} \circ \alpha$ followed by a translation $\tau_{\mathbf{w}}$.

Proposition 5.4. Every finite subgroup of isometries of n-dimensional space fixes at least one point.

Proof. Let G be a finite subgroup of isometries, and let $\mathbf{x}$ be any point of n-dimensional space. The orbit of $\mathbf{x}$ consists of a finite number of points that are permuted among themselves by any element of G. Since all the elements of G are rigid motions, the centroid of Orb $\mathbf{x}$ must always be sent to itself. Therefore, the centroid is a fixed point under G. □

If the fixed point of any finite subgroup G of isometries is taken as the origin, then G is a subgroup of $O(n)$, and all its elements can be written as orthogonal matrices. We now look at the structure of groups whose elements can be written as matrices.

MATRIX GROUPS

In physical sciences and in mathematical theory, we frequently encounter multiplicative group structures whose elements are $n \times n$ complex matrices. Such a group is called a **matrix group** if its identity element is the $n \times n$ identity matrix I. To investigate these groups, we have at our disposal, and shall freely apply, the machinery of linear algebra.

For example, if

$$A_k = \begin{pmatrix} \cos(2\pi k/m) & -\sin(2\pi k/m) \\ \sin(2\pi k/m) & \cos(2\pi k/m) \end{pmatrix},$$

then $(\{A_0, A_1, \ldots, A_{m-1}\}, \cdot)$ is a real matrix group of order m isomorphic to C_m. The matrix A_k represents a counterclockwise rotation of the plane about the origin through an angle $(2\pi k/m)$.

The matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}, \quad \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$
$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

form a group under matrix multiplication. This is a complex matrix group of order 8 that is, in fact, isomorphic to the quaternion group Q of Exercise 3.47.

Since the identity of any matrix group is the identity matrix I and every element of a matrix group must have an inverse, every element must be a nonsingular matrix. All the nonsingular $n \times n$ matrices over a field F form a group $(GL(n, F), \cdot)$ called the **general linear group** of dimension n over F. Any matrix group over the field F must be a subgroup of $GL(n, F)$.

Proposition 5.5. The determinant of any element of a finite matrix group must be an integral root of unity.

Proof. Let A be an element of a matrix group of order m. Then, by Corollary 4.10, $A^m = I$. Hence $(\det A)^m = \det A^m = \det I = 1$. □

Hence, if G is a real matrix group, the determinant of any element of G is either $+1$ or -1. If G is a complex matrix group, the determinant of any element is of the form $e^{2\pi ik/m}$.

The orthogonal group $O(n)$ is a real matrix group, and therefore any element must have determinant $+1$ or -1. The determinant function

$$\det: O(n) \to \{1, -1\}$$

is a group morphism from $(O(n), \cdot)$ to $(\{1, -1\}, \cdot)$. The kernel, consisting of orthogonal matrices with determinant $+1$, is called the **special orthogonal group** of dimension n and is denoted by

$$SO(n) = \{A \in O(n) | \det A = +1\}.$$

This is a normal subgroup of $O(n)$ of index 2. The elements of $SO(n)$ are called **proper rotations**, whereas the elements in the other coset of $O(n)$ by $SO(n)$, consisting of orthogonal matrices with determinant -1, are called **improper rotations**.

An $n \times n$ complex matrix A is called **unitary** if it is invertible and A^{-1} is the conjugate transpose of A. Thus the real unitary matrices are precisely the orthogonal matrices. Indeed, if $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \overline{\mathbf{y}}$ denotes the inner product in $\mathbb{C}^n$, the matrix A is unitary if and only if it preserves inner products in $\mathbb{C}^n$ (that is, $\langle A\mathbf{x}, A\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$), if and only if the columns of A are orthonormal.

The **unitary group** of dimension n, $U(n)$, consists of all $n \times n$ complex unitary matrices under multiplication. The **special unitary group**, $SU(n)$, is the subgroup of $U(n)$ consisting of those matrices with determinant $+1$. The group SU(3) received some publicity in 1964 when the Brookhaven National Laboratory discovered the fundamental particle called the omega-minus baryon. The existence and properties of this particle had been predicted by a theory that used SU(3) as a symmetry group of elementary particles.

Proposition 5.6. If $\lambda \in \mathbb{C}$ is an eigenvalue of any unitary matrix, then $|\lambda| = 1$.

Proof. Let λ be an eigenvalue and $\mathbf{x}$ a corresponding nonzero eigenvector of the unitary matrix A. Then $A\mathbf{x} = \lambda\mathbf{x}$, and since A preserves distances, $\|\mathbf{x}\| = \|A\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$. Since $\mathbf{x}$ is nonzero, it follows that $|\lambda| = 1$. □

The group $\{A_k | k = 0, 1, \ldots, m-1\}$ of rotations of the plane is a subgroup of SO(2), and the eigenvalues that occur are $e^{\pm(2\pi ik/m)}$. The matrix group isomorphic to the quaternion group Q is a subgroup of SU(2), and the eigenvalues that occur are ± 1 and $\pm i$.

Cayley's theorem (Theorem 3.38) showed that any group could be represented by a group of permutations. Another way to represent groups is by means of matrices. A **matrix representation** of a group G is a group morphism $\phi: G \to GL(n, F)$. This is equivalent to an action of G on an n-dimensional vector space over the field F, by means of linear transformations. The representation is called **faithful** if the kernel of ϕ is the identity. In this case, ϕ is injective and G is isomorphic to $\mathrm{Im}\phi$, a subgroup of the general linear group. Matrix representations provide powerful tools for studying groups because they lend themselves readily to calculation. As a result, most physical applications of group theory use representations.

It is possible to prove that any representation of a finite group over the real or complex field may be changed by a similarity transformation into a representation that uses only orthogonal or unitary matrices, respectively. Therefore, a real or complex faithful representation allows us to view a group as a subgroup of $\mathrm{O}(n)$ or $\mathrm{U}(n)$, respectively.

FINITE GROUPS IN TWO DIMENSIONS

We determine all the finite subgroups of rotations (proper and improper) of the plane $\mathbb{R}^2$. That is, we find all the finite matrix subgroups of SO(2) and O(2). This was essentially done by Leonardo da Vinci when he determined the possible symmetries of a central building with chapels attached. See Field and Golubitsky [29], where they construct interesting symmetric patterns in the plane using chaotic maps.

Proposition 5.7

(i) The set of proper rotations in two dimensions is

$$\mathrm{SO}(2) = \left\{ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \middle| \theta \in \mathbb{R} \right\}.$$

(ii) The set of improper rotations in two dimensions is

$$\left\{ \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \middle| \theta \in \mathbb{R} \right\}.$$

(iii) The eigenvalues of the proper rotation $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ are $e^{\pm i\theta}$ and those of any improper rotation are ± 1.

Proof. (i) Let $A = \begin{pmatrix} p & q \\ r & s \end{pmatrix} \in \mathrm{SO}(2)$, so that $A \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} p \\ r \end{pmatrix}$ and $A \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} q \\ s \end{pmatrix}$. Since A preserves distances, $p^2 + r^2 = 1$, and $q^2 + s^2 = 1$; thus there

exists angles θ and ϕ such that $p = \cos\theta, r = \sin\theta, q = \sin\phi$, and $s = \cos\phi$. Therefore,

$$\det A = ps - qr = \cos\theta\cos\phi - \sin\theta\sin\phi = \cos(\theta + \phi).$$

If A is proper, $\det A = 1$, so $\theta + \phi = 2n\pi$. Hence $\phi = 2n\pi - \theta$, and A is of the form $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. Conversely, if A is of this form, then $AA^T = I$ and $A \in \mathrm{O}(2)$. Since $\det A = +1$, A is a proper rotation, and $A \in \mathrm{SO}(2)$.

(ii) One improper rotation in $\mathbb{R}^2$ is $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, so the coset of improper rotations is

$$\mathrm{SO}(2)\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \left\{ \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \middle| \theta \in \mathbb{R} \right\}.$$

(iii) If λ is an eigenvalue of

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \text{then } \det\begin{pmatrix} (\cos\theta) - \lambda & -\sin\theta \\ \sin\theta & (\cos\theta) - \lambda \end{pmatrix} = 0.$$

Therefore, $\lambda^2 - 2\lambda\cos\theta + 1 = 0$ and $\lambda = \cos\theta \pm i\sin\theta = e^{\pm i\theta}$.

If λ is an eigenvalue of the improper rotation $\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$, then $\det\begin{pmatrix} (\cos\theta) - \lambda & \sin\theta \\ \sin\theta & -(\cos\theta) - \lambda \end{pmatrix} = 0$. Hence $\lambda^2 - 1 = 0$ and $\lambda = \pm 1$. □

The improper rotation $B = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ always has an eigenvalue 1 and hence leaves an axis through the origin invariant because, for any corresponding eigenvector $\mathbf{x}$, $B\mathbf{x} = \mathbf{x}$. It can be verified that this axis of eigenvectors, corresponding to the eigenvalue 1, is a line through the origin making an angle $\theta/2$ with the first coordinate axis. The matrix B corresponds to a reflection of the plane about this axis.

Hence we see that an improper rotation is a reflection about a line through the origin, and conversely, it is easy to see that a reflection about a line through the origin is an improper rotation.

Theorem 5.8. If G is a finite subgroup of SO(2), then G is cyclic, and so is isomorphic to C_n for some $n \in \mathbb{P}$.

Proof. By Proposition 5.6, every element $A \in G \subset \mathrm{SO}(2)$ is a counterclockwise rotation through an angle $\theta(A)$, where $0 \leqslant \theta(A) < 2\pi$. Since G is finite, we can choose an element $B \in G$ so that $\theta(B)$ is the smallest positive angle. For any $A \in G$, there exists an integer $r \geqslant 0$ such that $r\theta(B) \leqslant \theta(A) < (r+1)\theta(B)$. Since $\theta(AB^{-r}) = \theta(A) - r\theta(B)$, it follows that $0 \leqslant \theta(AB^{-r}) < \theta(B)$. Therefore, $\theta(AB^{-r}) = 0$, $AB^{-r} = I$, and $A = B^r$.

Hence $G = \{I, B, B^2, \ldots, B^r, \ldots, B^{n-1}\}$, and G is a finite cyclic group that must be isomorphic to C_n for some integer n. □

Theorem 5.9. If G is a finite subgroup of O(2), then G is isomorphic to either C_n or D_n for some $n \in \mathbb{P}$.

Proof. The kernel of the morphism det: $G \to \{1, -1\}$ is a normal subgroup, H, of index 1 or 2 consisting of the proper rotations in G. By the previous theorem, H is a cyclic group of order n, generated by B, for example.

If G contains no improper rotations, then $G = H \cong C_n$. If G does contain an improper rotation A, then

$$G = H \cup HA = \{I, B, B^2, \ldots, B^{n-1}, A, BA, B^2A, \ldots, B^{n-1}A\}.$$

Since A and B^kA are reflections, $A = A^{-1}$ and $B^kA = (B^kA)^{-1} = A^{-1}B^{-k} = AB^{n-k}$. These relations completely determine the multiplication in G, and it is now clear that G is isomorphic to the dihedral group D_n by an isomorphism that takes B to a rotation through $2\pi/n$ and A to a reflection. □

Theorem 5.9 shows that the only possible types of finite symmetries, fixing one point, of any geometric figure in the plane are the cyclic and dihedral groups. Examples of such symmetries abound in nature; the symmetry group of a snowflake is usually D_6, and many flowers have five petals with symmetry group C_5.

We have found all the possible *finite* symmetries in the plane that fix one point. However, there are figures in the plane that have *infinite* symmetry groups that fix one point; one example is the circular disk. The group of proper symmetries of this disk is the group SO(2), whereas the group of all symmetries is the whole of O(2).

PROPER ROTATIONS OF REGULAR SOLIDS

One class of symmetries that we know occurs in three dimensions is the class of symmetry groups of the regular solids: the tetrahedron, cube, octahedron, dodecahedron, and icosahedron. In this section, we determine the *proper* rotation groups of these solids. These will all be subgroups of SO(3). We restrict our consideration to proper rotations because these are the only ones that can be physically performed on models in three dimensions; to physically perform an improper symmetry on a solid, we would require four dimensions!

Theorem 5.10. Every element $A \in \mathrm{SO}(3)$ has a fixed axis, and A is a rotation about that axis.

Proof. Let λ_1, λ_2, and λ_3 be the eigenvalues of A. These are the roots of the cubic characteristic polynomial with real coefficients. Hence, at least one eigenvalue is real and if a second one is complex, the third is its complex conjugate.

By Proposition 5.6, $|\lambda_1| = |\lambda_2| = |\lambda_3| = 1$. Since $\det A = \lambda_1\lambda_2\lambda_3 = 1$, we can relabel the eigenvalues, if necessary, so that one of the following cases occurs:

(i) $\lambda_1 = \lambda_2 = \lambda_3 = 1$.
(ii) $\lambda_1 = 1, \lambda_2 = \lambda_3 = -1$.
(iii) $\lambda_1 = 1, \lambda_2 = \overline{\lambda}_3 = e^{i\theta}$ (where $\theta \neq n\pi$).

In all cases there is an eigenvalue equal to 1. If $\mathbf{x}$ is a corresponding eigenvector, then $A\mathbf{x} = \mathbf{x}$, and A fixes the axis along the vector $\mathbf{x}$. We can change the coordinate axes so that A can be written in one of the following three forms:

$$\text{(i)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{(ii)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \text{(iii)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}.$$

The first matrix is the identity, the second is a rotation through π, and the third is a rotation through θ about the fixed axis. □

A **regular solid** is a polyhedron in which all faces are congruent regular polygons and all vertices are incident with the same number of faces. There are five such solids, and they are illustrated in Figure 5.1; their structure is given in Table 5.1. The reader interested in making models of these polyhedra should consult Cundy and Rollett [28].

Given any polyhedron, we can construct its **dual** polyhedron in the following way. The vertices of the dual are the centers of the faces of the original polyhedron. Two centers are joined by an edge if the corresponding faces meet in

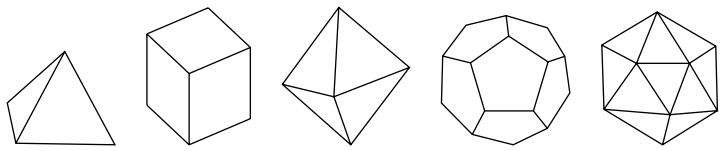


Figure 5.1. Regular solids.

TABLE 5.1. Regular Solids

Polyhedron	Number of Vertices	Number of Edges	Number of Faces	Faces	Number of Faces at Each Vertex
Tetrahedron	4	6	4	Triangles	3
Cube	8	12	6	Squares	3
Octahedron	6	12	8	Triangles	4
Dodecahedron	20	30	12	Pentagons	3
Icosahedron	12	30	20	Triangles	5

an edge. The dual of a regular tetrahedron is another regular tetrahedron. The dual of a cube is an octahedron, and the dual of an octahedron is a cube. The dodecahedron and icosahedron are also duals of each other. Any symmetry of a polyhedron will induce a symmetry on its dual and vice versa. Hence dual polyhedra will have the same rotation group.

Theorem 5.11. The group of proper rotations of a regular tetrahedron is isomorphic to A_4.

Proof. Label the vertices of the tetrahedron 1, 2, 3, and 4. Then any rotation of the tetrahedron will permute these vertices. So if G is the rotation group of the tetrahedron, we have a group morphism $f: G \to S_4$ whose kernel contains only the identity element. Hence, by the morphism theorem, G is isomorphic to $\mathrm{Im} f$.

We can use Theorem 4.40 to count the number of elements of G. The stabilizer of the vertex 1 is the set of elements fixing 1 and is $\{(1), (234), (243)\}$. The vertex 1 can be taken to any of the four vertices under G, so the orbit of 1 is the set of four vertices. Hence $|G| = |\mathrm{Stab}\ 1|\ |\mathrm{Orb}\ 1| = 3.4 = 12$.

There are two types of nontrivial elements in G that are illustrated in Figures 5.2 and 5.3. There are rotations of order 3 about axes, each of which joins a vertex to the center of the opposite face. These rotations perform an even permutation of the vertices because each fixes one vertex and permutes the other three cyclically. There are also rotations of order 2 about axes, each of which joins the midpoints of a pair of opposite edges. (Two edges in a tetrahedron are said to be opposite if they do not meet.) The corresponding permutations interchange two pairs of vertices and, being products of two transpositions, are even.

Hence $\mathrm{Im} f$ consists of 12 permutations, all of which are even, and $\mathrm{Im} f = A_4$. □

The alternating group A_4 is sometimes called the **tetrahedral group**.

There are many different ways of counting the number of elements of the rotation group G of the tetrahedron. One other way is as follows. Consider the tetrahedron sitting on a table, and shade in an equilateral triangle on the table where the bottom face rests, as in Figure 5.4. Any symmetry in G can be performed by picking up the tetrahedron, turning it, and replacing it on the table so that one face of the tetrahedron lies on top of the shaded equilateral triangle. Any of the *four* faces of the tetrahedron can be placed on the table, and each face can be placed on top of the shaded triangle in *three* different ways.

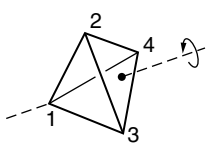


Figure 5.2. Element (2 3 4).

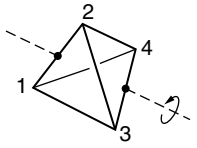


Figure 5.3. Element (1 2) ∘ (3 4).

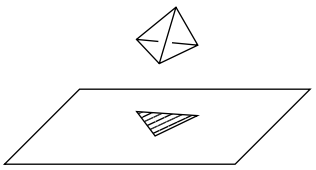


Figure 5.4

Hence $|G| = 4 \cdot 3 = 12$. This really corresponds to applying Theorem 4.40 to the stabilizer and orbit of a *face* of the tetrahedron.

Theorem 5.12. The group of proper rotations of a regular octahedron and cube is isomorphic to S_4.

Proof. The regular octahedron is dual to the cube, so it has the same rotation group. There are four diagonals in a cube that join opposite vertices. Label these diagonals 1, 2, 3, and 4 as in Figure 5.5. Any rotation of the cube will permute these diagonals, and this defines a group morphism $f: G \to S_4$, where G is the rotation group of the cube.

The stabilizer of any vertex of the cube is a cyclic group of order 3 that permutes the three adjacent vertices. The orbit of any vertex is the set of eight vertices. Hence, by Theorem 4.40, $|G| = 3 \cdot 8 = 24$.

Consider the rotation of order 2 about the line joining A to A' in Figure 5.5. The corresponding permutation is the transposition (12). Similarly, any other transposition is in $\mathrm{Im} f$. Therefore, by Proposition 3.35, $\mathrm{Im} f = S_4$.

By the morphism theorem, $G/\mathrm{Ker} f \cong S_4$ and $|G|/|\mathrm{Ker} f| = |S_4|$. Since $|G| = |S_4| = 24$, it follows that $|\mathrm{Ker} f| = 1$, and f is an isomorphism. □

The symmetric group S_4 is sometimes called the **octahedral group**.

Theorem 5.13. The group of proper rotations of a regular dodecahedron and a regular icosahedron is isomorphic to A_5.

Proof. A regular dodecahedron is dual to the icosahedron, so it has the same rotation group.

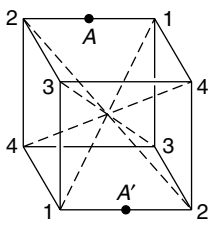


Figure 5.5. Diagonals of the cube.

There are 30 edges of an icosahedron, and there are 15 lines through the center joining the midpoints of opposite edges. (The reflection of each edge in the center of the icosahedron is a parallel edge, called the opposite edge.) Given any one of these 15 lines, there are exactly two others that are perpendicular both to the first line and to each other. We call three such mutually perpendicular lines a triad. The 15 lines fall into five sets of triads. Label these triads 1, 2, 3, 4, and 5. Figure 5.6 shows the top half of an icosahedron, where we have labeled the endpoints of each triad. (The existence of mutually perpendicular triads and the labeling of the diagram can best be seen by actually handling a model of the icosahedron.)

A rotation of the icosahedron permutes the five triads among themselves, and this defines a group morphism $f: G \to S_5$, where G is the rotation group of the icosahedron.

The stabilizer of any vertex of the icosahedron is a group of order 5 that cyclically permutes the five adjacent vertices. The orbit of any vertex is the set of all 12 vertices. Hence, by Theorem 4.40, $|G| = 5 \cdot 12 = 60$.

There are three types of nontrivial elements in G. There are rotations of order 5 about axes through a vertex. The rotations about the vertex A in Figure 5.6 correspond to multiples of the cyclic permutation (12345), all of which are even. There are rotations of order 3 about axes through the center of a face. The rotations about an axis through the point B, in Figure 5.6, are multiples of (142) and are therefore even permutations. Finally, there are rotations of order 2 about the 15 lines joining midpoints of opposite edges. The permutation corresponding to a rotation about an axis through C, in Figure 5.6, is $(23) \circ (45)$, which is even.

Every 3-cycle occurs in the image of f so, by Proposition 3.37, $\mathrm{Im} f = A_5$. Since G and A_5 both have 60 elements, the morphism theorem implies that G is isomorphic to A_5. □

The alternating group A_5 is sometimes called the **icosahedral group**.

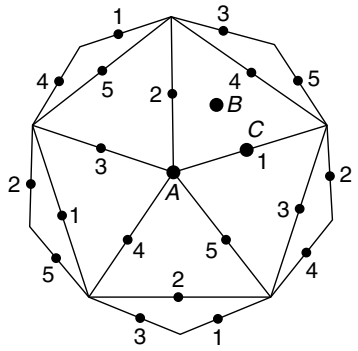


Figure 5.6. Ends of the triads of the icosahedron.

FINITE ROTATION GROUPS IN THREE DIMENSIONS

We now proceed to show that the only finite proper rotation groups in three dimensions are the three symmetry groups of the regular solids, A_4, S_4, and A_5 together with the cyclic and dihedral groups, C_n and D_n.

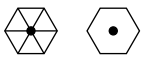
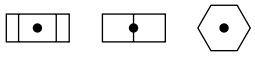
The unit sphere $S^2 = \{\mathbf{x} \in \mathbb{R}^3 \mid \|\mathbf{x}\| = 1\}$ is mapped to itself by every element of O(3). Every rotation group fixing the origin is determined by its action on the unit sphere S^2. By Theorem 5.10, every nonidentity element $A \in \mathrm{SO}(3)$ leaves precisely two antipodal points on S^2 fixed. That is, there exists $\mathbf{x} \in S^2$ such that $A(\mathbf{x}) = \mathbf{x}$ and $A(-\mathbf{x}) = -\mathbf{x}$. The points $\mathbf{x}$ and $-\mathbf{x}$ are called the **poles** of A. Let P be the set of poles of the nonidentity elements of a finite subgroup G of SO(3).

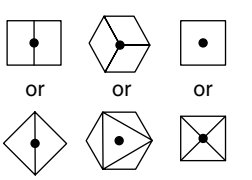
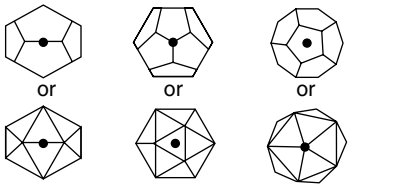
Proposition 5.14. *G acts on the set, P, of poles of its nonidentity elements.*

Proof. We show that G permutes the poles among themselves. Let A, B, be nonidentity elements of G, and let $\mathbf{x}$ be a pole of A. Then $(BAB^{-1})B(\mathbf{x}) = BA(\mathbf{x}) = B(\mathbf{x})$, so that $B(\mathbf{x})$ is a pole of BAB^{-1}. Therefore, the image of any pole is another pole, and G acts on the set of poles. □

We classify the rotation groups by considering the number of elements in the stabilizers and orbits of the poles. Recall that the stabilizer of a pole $\mathbf{x}$, $\mathrm{Stab}\,\mathbf{x} = \{A \in G \mid A(\mathbf{x}) = \mathbf{x}\}$, is a subgroup of G, and that the orbit of $\mathbf{x}$, $\mathrm{Orb}\,\mathbf{x} = \{B(\mathbf{x}) \mid B \in G\}$, is a subset of the set P of poles. In Table 5.2 we look at the stabilizers and orbits of the poles of the rotation groups we have already discussed.

TABLE 5.2. Poles of the Finite Rotation Groups

Group G =	C_n	D_n	A_4
$\lvert G\rvert$ =	n	$2n$	12
Symmetries of	n-agonal cone	n-agonal cylinder	tetrahedron
Looking down on the pole, $\mathbf{x}$			
$\lvert\mathrm{Stab}\,\mathbf{x}\rvert$ =	n n	2 2 n	2 3 3
$\lvert\mathrm{Orb}\,\mathbf{x}\rvert$ =	1 1	n n 2	6 4 4

Group G =	S_4	A_5
$\lvert G\rvert$ =	24	60
Symmetries of	cube or octahedron	dodecahedron or icosahedron
Looking down on the pole, $\mathbf{x}$		
$\lvert\mathrm{Stab}\,\mathbf{x}\rvert$ =	2 3 4	2 3 5
$\lvert\mathrm{Orb}\,\mathbf{x}\rvert$ =	12 8 6	30 20 12

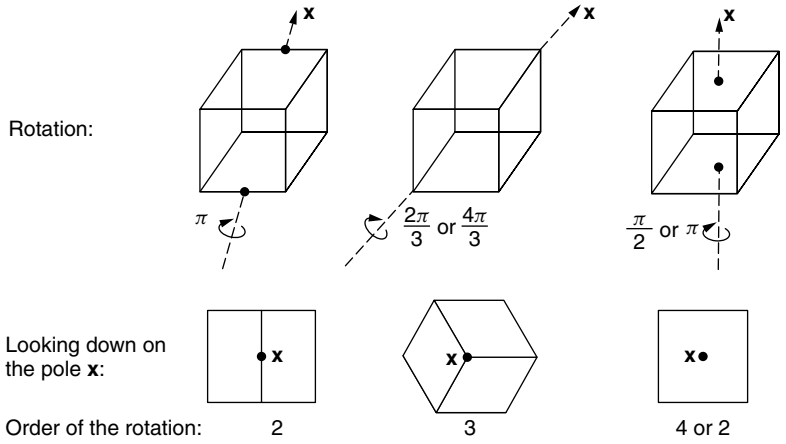


Figure 5.7. Rotations of the cube.

We take C_n to be the rotation group of a regular n-agonal cone whose base is a regular n-gon. (The sloping edges of the cone must not be equal to the base edges if $n = 3$.) D_n is the rotation group of a regular n-agonal cylinder whose base is a regular n-gon. (The vertical edges must not be equal to the base edges if $n = 4$.)

Each stabilizer group, Stab $\mathbf{x}$, is a cyclic subgroup of rotations of the solid about the axis through $\mathbf{x}$. The orbit of $\mathbf{x}$, Orb $\mathbf{x}$, is the set of poles of the same type as $\mathbf{x}$. As a check on the number of elements in the stabilizers and orbits, we have $|G| = |\text{Stab } \mathbf{x}||\text{Orb } \mathbf{x}|$ for each pole $\mathbf{x}$.

For example, the cube has three types of poles and four types of nontrivial elements in its rotation group; these are illustrated in Figure 5.7.

Theorem 5.15. Any finite subgroup of SO(3) is isomorphic to one of

$$C_n (n \geqslant 1),\ D_n (n \geqslant 2),\ A_4,\ S_4 \text{ or } A_5.$$

Proof. Let G be a finite subgroup of SO(3). Choose a set of poles $\mathbf{x}_1, \ldots, \mathbf{x}_r$, one from each orbit. Let $p_i = |\text{Stab } \mathbf{x}_i|$ and $q_i = |\text{Orb } \mathbf{x}_i|$, so that $p_i q_i = n = |G|$.

Each nonidentity element of G has two poles; thus the total number of poles, counting repetitions, is $2(n-1)$. The pole $\mathbf{x}_i$ occurs as a pole of a nonidentity element $p_i - 1$ times. There are q_i poles of the same type as $\mathbf{x}_i$. Therefore, the total number of poles, counting repetitions, is

$$2(n-1) = \sum_{i=1}^{r} q_i (p_i - 1) = \sum_{i=1}^{r} (n - q_i),$$

so

$$2 - \frac{2}{n} = \sum_{i=1}^{r} \left(1 - \frac{1}{p_i}\right). \qquad (*)$$

If G is not the trivial group, $n \geqslant 2$ and $1 \leqslant 2 - \dfrac{2}{n} < 2$. Since $\mathbf{x}_i$ is a pole of some nonidentity element, Stab $\mathbf{x}_i$ contains a nonidentity element, and $p_i \geqslant 2$. Therefore, $\dfrac{1}{2} \leqslant 1 - \dfrac{1}{p_i} < 1$. It follows from (*) that the number of orbits, r, must be 2 or 3.

If there are just two orbits, it follows that

$$2 - \frac{2}{n} = 1 - \frac{1}{p_1} + 1 - \frac{1}{p_2}$$

and $2 = \dfrac{n}{p_1} + \dfrac{n}{p_2} = q_1 + q_2$. Hence $q_1 = q_2 = 1$, and $p_1 = p_2 = n$. This means that $\mathbf{x}_1 = \mathbf{x}_2$, and there is just one axis of rotation. Therefore, G is isomorphic to the cyclic group C_n.

If there are three orbits, it follows that

$$2 - \frac{2}{n} = 1 - \frac{1}{p_1} + 1 - \frac{1}{p_2} + 1 - \frac{1}{p_3}$$

$$1 + \frac{2}{n} = \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3}. \qquad (**)$$

Suppose that $p_1 \leqslant p_2 \leqslant p_3$. If $p_1 \geqslant 3$, we would have

$$\frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} \leqslant \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1,$$

which is a contradiction, since $\dfrac{2}{n} > 0$. Hence $p_1 = 2$ and $q_1 = \dfrac{n}{2}$.

Now $1 + \dfrac{2}{n} = \dfrac{1}{2} + \dfrac{1}{p_2} + \dfrac{1}{p_3}$, so $\dfrac{1}{2} + \dfrac{2}{n} = \dfrac{1}{p_2} + \dfrac{1}{p_3}$. If $p_2 \geqslant 4$, we would have $\dfrac{1}{p_2} + \dfrac{1}{p_3} \leqslant \dfrac{1}{4} + \dfrac{1}{4} = \dfrac{1}{2}$, which is a contradiction, since $\dfrac{2}{n} > 0$. Hence p_2 is 2 or 3. The only possibilities are the following.

Case (i). $p_1 = 2$, $p_2 = 2$, $p_3 = \dfrac{n}{2}$, n is even and $n \geqslant 4$, $q_1 = \dfrac{n}{2}$, $q_2 = \dfrac{n}{2}$, and $q_3 = 2$.

Case (ii). $p_1 = 2$, $p_2 = 3$, $p_3 = 3$, $n = 12$, $q_1 = 6$, $q_2 = 4$, and $q_3 = 4$.

Case (iii). $p_1 = 2$, $p_2 = 3$, $p_3 = 4$, $n = 24$, $q_1 = 12$, $q_2 = 8$, and $q_3 = 6$.

Case (iv). $p_1 = 2$, $p_2 = 3$, $p_3 = 5$, $n = 60$, $q_1 = 30$, $q_2 = 20$, and $q_3 = 12$.

If $p_2 = 2$ and $p_3 \geqslant 6$, $\dfrac{1}{p_2} + \dfrac{1}{p_3} \leqslant \dfrac{1}{3} + \dfrac{1}{6} = \dfrac{1}{2}$, which contradicts (**), since $\dfrac{2}{n} > 0$.

Case (i). Let $H = \text{Stab } \mathbf{x}_3$. This is a group of rotations about one axis, and it is a cyclic group of order $n/2$. Any other element A that is not in H is of order 2 and is a half turn. Therefore, $G = H \cup HA$, and G is isomorphic to $D_{n/2}$ of order n.

Case (ii). Let $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$, and $\mathbf{y}_4$ be the four poles in Orb $\mathbf{x}_2$. Now $p_2 = |\text{Stab } \mathbf{y}_i| = 3$; thus Stab $\mathbf{y}_1$ permutes $\mathbf{y}_2$, $\mathbf{y}_3$, and $\mathbf{y}_4$ as in Figure 5.8. Therefore, $\|\mathbf{y}_2 - \mathbf{y}_1\| = \|\mathbf{y}_3 - \mathbf{y}_1\| = \|\mathbf{y}_4 - \mathbf{y}_1\|$. We have similar results for Stab $\mathbf{y}_2$ and Stab $\mathbf{y}_3$. Hence $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$, and $\mathbf{y}_4$ are the vertices of a regular tetrahedron, and G is a subgroup of the symmetries of this tetrahedron. Since $|G| = 12$, G must be the whole rotation group, A_4.

Case (iii). Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_6$ be the six poles in Orb $\mathbf{x}_3$. Since $p_3 = 4$, a rotation in Stab $\mathbf{y}_i$ must fix two of the poles and rotate the other four cyclically. Hence $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_6$ must lie at the vertices of a regular octahedron. Again, since $|G| = 24$, G must be the whole rotation group, S_4, of this octahedron.

Case (iv). Let $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{12}$ be the 12 poles in Orb $\mathbf{x}_3$. Any element of order 5 in G must permute these poles and hence must fix two poles and permute the others, as in Figure 5.9, in two disjoint 5-cycles, say $(2\ \ 3\ \ 4\ \ 5\ \ 6) \circ (7\ \ 8\ \ 9\ \ 10\ \ 11)$, where we denote the pole $\mathbf{y}_i$ by i. The points $\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5$, and $\mathbf{y}_6$ form a regular pentagon and their distances from $\mathbf{y}_1$ are all equal. Using similar results for rotations of order 5 about the other poles, we see that the poles are the vertices of an icosahedron, and the group G is the proper rotation group, A_5, of this icosahedron. □

Throughout this section we have considered only *proper* rotations. However, if we allow *improper* rotations as well, it can be shown that a finite subgroup

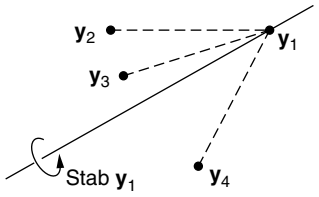


Figure 5.8

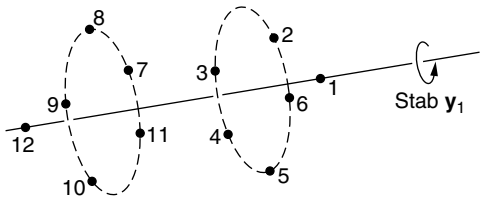


Figure 5.9

of O(3) is isomorphic to one of the groups in Theorem 5.15 or contains one of these groups as a normal subgroup of index 2. See Coxeter [27, Sec. 15.5] for a more complete description of these improper rotation groups.

CRYSTALLOGRAPHIC GROUPS

This classification of finite symmetries in $\mathbb{R}^3$ has important applications in crystallography. Many chemical substances form crystals and their structures take the forms of crystalline lattices. A crystal lattice is always finite, but in order to study its symmetries, we create a mathematical model by extending this crystal lattice to infinity. We define an *ideal crystalline lattice* to be an infinite set of points in $\mathbb{R}^3$ of the form

$$n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3,$$

where $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ form a basis of $\mathbb{R}^3$ and $n_1, n_2, n_3 \in \mathbb{Z}$. Common salt forms a cubic crystalline lattice in which $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ are orthogonal vectors of the same length. Figure 5.10 illustrates a crystalline lattice.

This use of the term *lattice* is not the same as that in Chapter 2, where a lattice referred to a special kind of partially ordered set. To avoid confusion, we always use the term *crystalline lattice* here.

A subgroup of O(3) that leaves a crystalline lattice invariant is called a **crystallographic point group**. This is a finite subgroup of O(3) because there are only a finite number of crystalline lattice points that can be the images of $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ when the origin is fixed.

However, not all finite subgroups of O(3) are crystallographic point groups. Suppose that $A \in \mathrm{SO}(3)$ leaves a crystalline lattice $\mathcal{L}$ invariant. Then, by Theorem 5.10, A is a rotation through an angle θ, and the trace of A is $1 + 2\cos\theta$. If we choose a basis for $\mathbb{R}^3$ consisting of the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ of the crystalline lattice $\mathcal{L}$, the matrix representing A will have integer entries. The trace is invariant under change of basis, so the trace of A must be an integer. Hence $2\cos\theta$

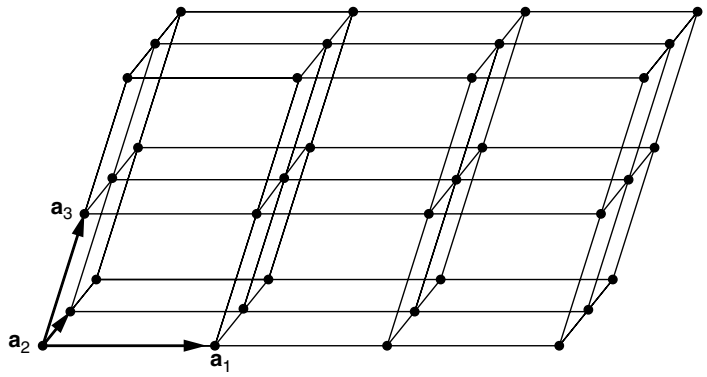


Figure 5.10. Crystalline lattice.

must be integral, and θ must be either $k\pi/2$ or $k\pi/3$, where $k \in \mathbb{Z}$. It follows that every element of a crystallographic point group in SO(3) can only contain elements of order 1, 2, 3, 4, or 6.

It can be shown that every crystallographic point group in SO(3) is isomorphic to one of C_1, C_2, C_3, C_4, C_6, D_2, D_3, D_4, D_6, A_4, or S_4.

If we allow reflections, the only other such groups in O(3) must contain one of these groups as a normal subgroup of index 2. Every one of these groups occurs in nature as the point group of at least one chemical crystal. See Coxeter [27, Sec. 15.6] or Lomont [31, Chap. 4, Sec. 4].

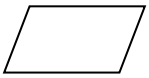
EXERCISES

Find the group of proper rotations and the group of all rotations of the figures in Exercises 5.1 to 5.7.

5.1. 

5.2. 

5.3. 

5.4. 

5.5. 

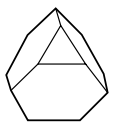
5.6. 

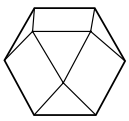
5.7. 

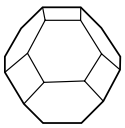
5.8. Let G be the subgroup of O(2) isomorphic to D_n. Find two matrices A and B so that any element of G can be written as a product of A's and B's.

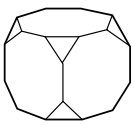
5.9. What is the group of proper rotations of a rectangular box of length 3 cm, depth 2 cm, and height 2 cm?

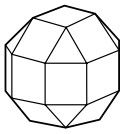
Find the proper rotation group of the 13 **Archimedean solids** *in Exercises 5.10 to 5.22. All the faces of these solids are regular polygons and all the vertices are similar. (See Cundy and Rollet [28] for methods on how to construct these solids.)*

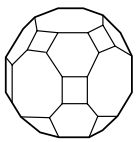
5.10. 

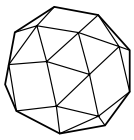
5.11. 

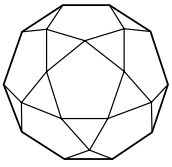
5.12. 

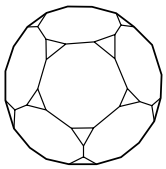
5.13. 

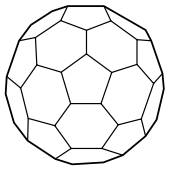
5.14. 

5.15. 

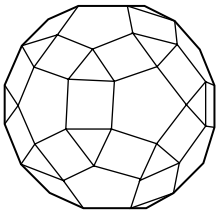
5.16. 

5.17. 

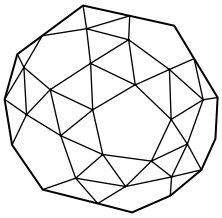
5.18. 

5.19. 

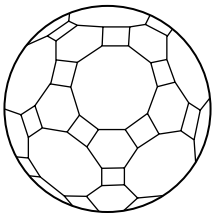
5.20.



5.21.



5.22.



5.23. It is possible to inscribe five cubes in a regular dodecahedron. One such cube is shown in Figure 5.11. Use these cubes to show that the rotation group of the dodecahedron is A_5.

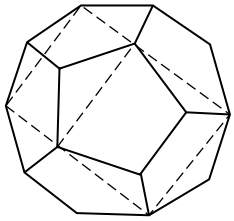


Figure 5.11. Cube inside a dodecahedron.

5.24. What is the proper symmetry group of a cube in which three faces, coming together at one vertex, are painted green and the other faces are red?

5.25. Find the group of all rotations (both proper and improper) of a regular tetrahedron.

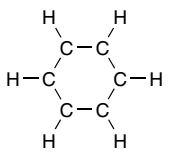
5.26. Let G be the full symmetry group of the cube. Define $f: G \to S_4$ as in Theorem 5.12. Find the kernel of f and the order of G.

5.27. Let the vertices of a tetrahedron be $(1, 1, 1)$, $(-1, -1, 1)$, $(-1, 1, -1)$, and $(1, -1, -1)$. Find matrices in SO(3) of orders 2 and 3 that leave the tetrahedron invariant.

5.28. Let the vertices of a cube be $(\pm 1, \pm 1, \pm 1)$. Find matrices in SO(3) of orders 2, 3, and 4 that leave the cube invariant.

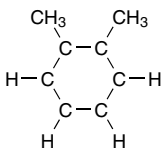
Find the symmetry groups of the chemical molecules in Exercises 5.29 to 5.31. (Assume that all of the C-C bonds are equivalent.)

5.29.



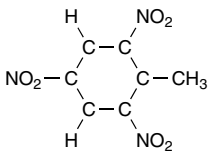
Benzene

5.30.



Xylene

5.31.



Trinitrotoluene (TNT)

Find matrices in SO(3) that preserve the crystalline lattices described in Exercises 5.32 to 5.34, and find their crystallographic point groups. The points of the crystalline lattice are $n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3$, *where* $n_i \in \mathbb{Z}$ *and the basis vectors* $\mathbf{a}_i$ *are given below.*

5.32. $\mathbf{a}_1 = (1, 1, 0)$, $\mathbf{a}_2 = (-1, 1, 0)$, $\mathbf{a}_3 = (0, 1, 1)$.

5.33. $\mathbf{a}_1 = (1, 0, 0)$, $\mathbf{a}_2 = (0, 1, 0)$, $\mathbf{a}_3 = (0, 0, 2)$.

5.34. $\mathbf{a}_1 = (1, 0, 0)$, $\mathbf{a}_2 = (0, -3\sqrt{3}, 3)$, $\mathbf{a}_3 = (0, 3\sqrt{3}, 3)$.

5.35. Let $G(n)$ denote the group of linear isometries of $\mathbb{R}^n$.

(**a**) Show that $E(n) = T(n)G(n) = G(n)T(n)$, where the product of subgroups is as defined in Exercise 4.43.

(**b**) Show that $T(n) \cap G(n) = \{1_{\mathbb{R}^n}\}$.

(**c**) Use parts (a) and (b) to prove that $E(n)/T(n) \cong G(n)$.

6

PÓLYA–BURNSIDE METHOD OF ENUMERATION

This chapter provides an introduction to the Pólya–Burnside method of counting the number of orbits of a set under the action of a symmetry group. If a group G acts on a set X and we know the number of elements of X, this method will enable us to count the number of different types of elements of X under the action of G.

For example, how many different chemical compounds can be obtained by attaching a CH_3 or H radical to each carbon atom in the benzene ring of Figure 6.3? There are 2^6 different ways of attaching a CH_3 or H radical on paper, but these do not all give rise to different compounds because many are equivalent under a symmetry. There are six different ways of attaching one CH_3 radical and five H radicals, but they all give rise to the same compound. The dihedral group D_6 acts on the 2^6 ways of attaching the radicals, and the number of different compounds is the number of orbits under the action of D_6, that is, the number of formulas that cannot be obtained from each other by any rotation or reflection.

We have seen that the number of different switching circuits that can be obtained with n switches is 2^{2^n}. This number grows very quickly as n becomes large. Table 2.11 gives the 16 switching functions of two variables; when $n = 3$, there are 256 different circuits, and when $n = 4$, there are 65,536 different circuits. However, many of these circuits are equivalent if we change the labels of the switches. That is, the symmetric group, S_n, acts on the 2^{2^n} different circuits by permuting the labels of the switches. The number of nonequivalent circuits is the number of orbits under the action of S_n.

BURNSIDE'S THEOREM

Let G be a finite group that acts on a finite set X. The following theorem describes the number of orbits in terms of the number of elements left fixed

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

by each element of G. It was first proved by W. Burnside in 1911 and was called *Burnside's lemma*; it was not until 1937 that its applicability to many combinatorial problems was discovered by G. Pólya.

Theorem 6.1. Burnside's Theorem. Let G be a finite group that acts on the elements of a finite set X. For each $g \in G$, let Fix $g = \{x \in X | g(x) = x\}$, the set of elements of X left fixed by g. If N is the number of orbits of X under G, then

$$N = \frac{1}{|G|} \sum_{g \in G} |\text{ Fix } g|.$$

Proof. We count the set $S = \{(x, g) \in X \times G | g(x) = x\}$ in two different ways. Consider Table 6.1, whose columns are indexed by the elements of X and whose rows are indexed by the elements of G. Put a value of 1 in the (x, g) position if $g(x) = x$; otherwise, let the entry be 0.

The sum of the entries in row g is the number |Fix g| of elements left fixed by g. The sum of the entries in column x is |Stab x|, the number of elements of G that fix x.

We can count the number of elements of S either by totaling the row sums or by totaling the column sums. Hence

$$|S| = \sum_{g \in G} |\text{Fix } g| = \sum_{x \in X} |\text{Stab } x|.$$

Choose a set of representatives, $x_1, x_2, \ldots, x_N$, one from each orbit of X under G. If x is in the same orbit as x_i, then Orb x = Orb x_i, and by Theorem 4.40, |Stab x| = |Stab x_i|. Hence

$$\sum_{g \in G} |\text{Fix } g| = \sum_{i=1}^{N} \sum_{x \in \text{Orb } x_i}^{\cdot} |\text{Stab } x| = \sum_{i=1}^{N} |\text{Orb } x_i||\text{Stab } x_i| = N \cdot |G|$$

by Theorem 4.40. The theorem now follows. □

TABLE 6.1. Elements of S Correspond to the 1's in This Table

		Elements of X						Row Sums
					x			↓
					⋮			
Elements of G	g	…1	0	1	1	1	0…	|Fix g|
						0		
					⋮			
	Column → Sums				|Stab x|			

NECKLACE PROBLEMS

Example 6.2. Three black and six white beads are strung onto a circular wire. This necklace can be rotated in its plane, and turned over. How many different types of necklaces can be made assuming that beads of the same color are indistinguishable?

Solution. Position the three black and six white beads at the vertices of a regular 9-gon. If the 9-gon is fixed, there are $9 \cdot 8 \cdot 7/3! = 84$ different ways of doing this. Two such arrangements are equivalent if there is an element of the symmetry group of the regular 9-gon, D_9, which takes one arrangement into the other. The group D_9 permutes the different arrangements, and the number of nonequivalent arrangements is the number, N, of orbits under D_9. We can now use the Burnside theorem to find N.

Table 6.2 lists all the different types of elements of D_9 and the number of fixed points for each type. For example, consider the reflection $g \in D_9$ about the line joining vertex 2 to the center of the circle, which is illustrated in Figure 6.1. Then the arrangements that are fixed under g occur when the black beads are at vertices 1 2 3, 9 2 4, 8 2 5, or 7 2 6. Hence $|\text{Fix } g| = 4$. There are nine reflections about a line, one through each vertex. Therefore, the total number of fixed points contributed by these types of elements is $9 \cdot 4 = 36$. A rotation of order 3 in D_9 will fix an arrangement if the black beads are at vertices 1 4 7, 2 5 8, or 3 6 9; hence there are three arrangements that are fixed. If an arrangement is fixed under a rotation of order 9, all the beads must be the same color; hence $|\text{Fix } g| = 0$ if g has order 9. Table 6.2 shows that the sum of all the numbers of fixed points is 126. By Theorem 6.1, $N = \dfrac{1}{|D_9|} \sum_{g \in D_9} |\text{Fix } g| = \dfrac{126}{18} = 7$, and there are seven different types of necklaces. □

In this example, it is easy to determine all the seven types. They are illustrated in Figure 6.2.

TABLE 6.2. Action of D_9 on the Necklaces

Type of Element, $g \in D_9$	Order of g	Number, s, of Such Elements	$\lvert\text{Fix } g\rvert$	$s \cdot \lvert\text{Fix } g\rvert$
Identity	1	1	84	84
Reflection about a line	2	9	4	36
Rotation through $2\pi/3$ or $4\pi/3$	3	2	3	6
Other rotations	9	6	0	0
				$\sum = 126$

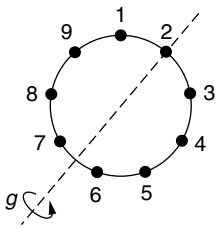


Figure 6.1. D_9 acting on the necklace.

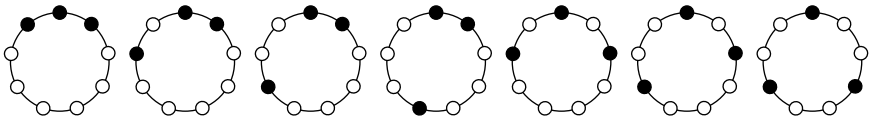


Figure 6.2. Seven types of necklaces.

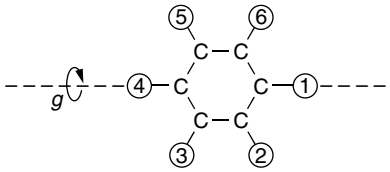


Figure 6.3. Benzene ring.

TABLE 6.3. Action of D_6 on the Compounds

Type of Element, $g \in D_6$	Order of g	Number, s, of Such Elements	$\lvert\text{Fix } g\rvert$	$s \cdot \lvert\text{Fix } g\rvert$
Identity	1	1	2^6	64
Reflection in a line through opposite vertices [e.g., $(26)\circ(35)\circ(1)\circ(4)$]	2	3	2^4	48
Reflection in a line through midpoints of opposite sides [e.g., $(56)\circ(14)\circ(23)$]	2	3	2^3	24
Rotation through $\pm\pi/3$ [e.g., (123456)]	6	2	2	4
Rotation through $\pm 2\pi/3$ [e.g., $(135)\circ(246)$]	3	2	2^2	8
Rotation through π, $(14)\circ(25)\circ(36)$	2	1	2^3	8
		$\lvert D_6\rvert = 12$		$\sum = 156$

Example 6.3. Find the number of different chemical compounds that can be obtained by attaching CH_3 or H radicals to a benzene ring.

Solution. The carbon atoms are placed at the six vertices of a regular hexagon, and there are 2^6 ways of attaching CH_3 or H radicals. The dihedral group D_6 acts on these 2^6 ways, and we wish to find the number of orbits.

Consider a reflection, g, about a line through opposite vertices. The order of g is 2, and there are three such reflections, through the three opposite pairs of vertices. |Fix g| can be determined by looking at Figure 6.3. If a configuration is fixed by g, the radical in place 2 must be the same as the radical in place 6, and also the radicals in places 3 and 5 must be equal. Hence the radicals in places 1, 2, 3, and 4 can be chosen arbitrarily, and this can be done in 2^4 ways.

The number of configurations left fixed by each element of D_6 is given in Table 6.3. To check that we have not omitted any elements, we add the column containing the numbers of elements, and this should equal the order of the group D_6. It follows from the Burnside theorem that the number of orbits is $156/|D_6| = 156/12 = 13$. Hence there are 13 different types of molecules obtainable. □

COLORING POLYHEDRA

Example 6.4. How many ways is it possible to color the vertices of a cube if n colors are available?

Solution. If the cube is fixed, the eight vertices can each be colored in n ways, giving a total of n^8 colorings. The rotation group of the cube, S_4, permutes these colorings among themselves, and the number of orbits is the number of distinct colorings taking the rotations into account. We can calculate the number of orbits using the Burnside theorem.

There are five types of elements in the rotation group of the cube. We take an element g of each type and determine the vertices that must have the same color in order that the coloring be invariant under g.

Figure 6.4 illustrates the different types of rotations; vertices that have to have the same color are shaded in the same way. Table 6.4 gives the number of fixed colorings, with column totals.

By the Burnside theorem, the number of orbits and hence the number of colorings is

$$\frac{1}{|S_4|}\sum_{g\in S_4}|\text{Fix } g| = \tfrac{1}{24}(n^8 + 17n^4 + 6n^2).$$

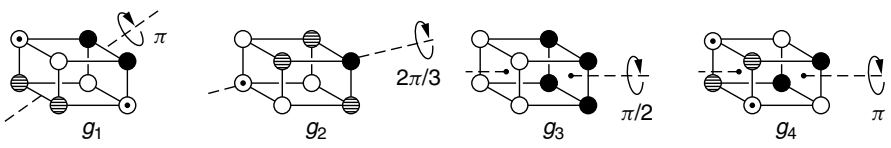


Figure 6.4. Types of rotations of the cube.

TABLE 6.4. Colorings of the Vertices of the Cube

Type of Element, g_i (Figure 6.4)	Order of g_i	Number, s, of Such Elements	Number, r, of Choices of Colors	$\lvert\text{Fix } g_i\rvert$ Which Is n^r	$s \cdot \lvert\text{Fix } g_i\rvert$
Identity	1	1	8	n^8	n^8
g_1	2	$6 \cdot 1 = 6$	4	n^4	$6n^4$
g_2	3	$4 \cdot 2 = 8$	4	n^4	$8n^4$
g_3	4	$3 \cdot 2 = 6$	2	n^2	$6n^2$
g_4	2	$3 \cdot 1 = 3$	4	n^4	$3n^4$
		$\lvert S_4\rvert = 24$			$n^8 + 17n^4 + 6n^2$

This shows, incidentally, that $n^8 + 17n^4 + 6n^2$ is divisible by 24 for all $n \in \mathbb{P}$. □

Example 6.5. In how many ways is it possible to color a regular dodecahedron so that five of its faces are black and the other seven are white?

Solution. The number of ways[†] of choosing five faces of a fixed dodecahedron to be colored black is $\binom{12}{5} = \dfrac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{5!} = 792.$

The different types of elements in the rotation group, A_5, of the dodecahedron are shown in Figure 6.5. The numbers of elements of a given type, in Table 6.5, are calculated as follows. An element of order 3 is a rotation about an axis through opposite vertices. Since there are 20 vertices, there are ten such axes. There are two nonidentity rotations of order 3 about each axis; thus the total number of elements of order 3 is $10 \cdot 2 = 20$. The elements of orders 2 and 5 can be counted in a similar way.

If $g_2 \in A_5$ is of order 3, we can calculate $\lvert\text{Fix } g_2\rvert$ as follows. The element g_2 does not fix any face and permutes the faces in disjoint 3-cycles. Now five black faces cannot be permuted by disjoint 3-cycles without fixing two faces, so $\lvert\text{Fix } g_2\rvert = 0$. Similarly, $\lvert\text{Fix } g_1\rvert = 0$ if g_1 is a 2-cycle. If g_3 is of order 5, then g_3 is a rotation about an axis through the centers of two opposite faces, and these two faces are fixed. The other ten faces are permuted in two disjoint 5-cycles; either of these 5-cycles can be black; thus $\lvert\text{Fix } g_3\rvert = 2$.

It follows from Table 6.5 and from the Burnside theorem that the number of different colorings is $840/60 = 14$. □

Any face coloring of the dodecahedron corresponds to a vertex coloring of its dual, the icosahedron.

[†] A set of n elements has $\binom{n}{k}$ subsets of $k \leqslant n$ elements where $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$ is the **binomial coefficient**.

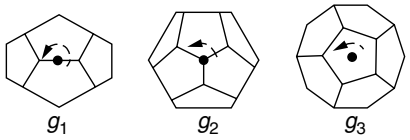


Figure 6.5. Types of rotations of a dodecahedron.

TABLE 6.5. Colorings of the Dodecahedron

Type of Element, g_i (Figure 6.5)	Order of g_i	Number, s, of Such Elements	$\lvert\text{Fix } g_i\rvert$	$s \cdot \lvert\text{Fix } g_i\rvert$
Identity	1	1	792	792
g_1	2	15	0	0
g_2	3	$10 \cdot 2 = 20$	0	0
g_3	5	$6 \cdot 4 = 24$	2	48
		$\lvert A_5\rvert = 60$		$\sum = 840$

COUNTING SWITCHING CIRCUITS

The Burnside theorem can still be applied when the sets to be enumerated do not have any geometric symmetry. In this case, the symmetry group is usually the full permutation group S_n.

Consider the different switching circuits obtained by using three switches. We can think of these as black boxes with three binary inputs x_1, x_2, and x_3 and one binary output $f(x_1, x_2, x_3)$, as in Figure 6.6. Two circuits, f and g, are called *equivalent* if there is a permutation π of the variables so that $f(x_1, x_2, x_3) = g(x_{\pi 1}, x_{\pi 2}, x_{\pi 3})$. Equivalent circuits can be obtained from each other by just permuting the wires outside the black boxes, as in Figure 6.7.

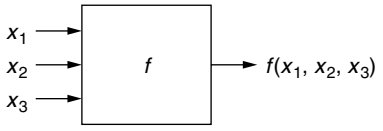


Figure 6.6. Switching circuit.

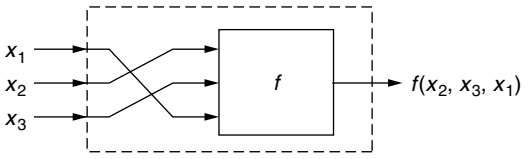


Figure 6.7. Permutation of the inputs.

Example 6.6. Find the number of switching circuits using three switches that are not equivalent under permutations of the inputs.

Solution. There are eight possible inputs using three binary variables and hence there are $2^8 = 256$ circuits to consider. The symmetric group S_3 acts on these 256 circuits, and we wish to find the number of different equivalence classes, that is, the number of orbits.

Table 6.6 lists the number of circuits left fixed by the different types of elements in S_3. For example, if the switching function $f(x_1, x_2, x_3)$ is fixed by the transposition (12) of the input variables, then $f(0, 1, 0) = f(1, 0, 0)$ and $f(0, 1, 1) = f(1, 0, 1)$. The values of f for the inputs (0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 1, 0), and (1, 1, 1) can be chosen arbitrarily in 2^6 ways.

By Burnside's theorem and Table 6.6, the number of nonequivalent circuits is $480/|S_3| = 480/6 = 80$. □

However, this number can be reduced further if we allow permutations and complementation of the three variables. In a circuit consisting of two-state switches, the variable x_i can be complemented by simply reversing each of the switches controlled by x_i. The resulting circuit is just as simple and the cost is the same as the original one. In transistor networks, we can just permute the input wires and add NOT gates as in Figure 6.8.

The eight input values of a three-variable switching circuit can be considered as the vertices of a three-dimensional cube, as shown in Figure 6.9. The six faces of this cube are defined by the equations $x_1 = 0$, $x_1 = 1$, $x_2 = 0$, $x_2 = 1$, $x_3 = 0$, $x_3 = 1$. The group that permutes and complements the variables takes each face to another face and takes opposite faces to opposite faces. Hence the group is the complete symmetry group, G, of the cube. There is a morphism $\psi: G \to \{1, -1\}$,

TABLE 6.6. Action of S_3 on the Inputs of the Switches

Type of Element, $g \in S_3$	Number, s, of Such Elements	$\lvert\text{Fix } g\rvert$	$s \cdot \lvert\text{Fix } g\rvert$
Identity	1	2^8	$2^8 = 256$
Transposition	3	2^6	$3 \cdot 2^6 = 192$
3-Cycle	2	2^4	$2 \cdot 2^4 = 32$
	$\lvert S_3\rvert = 6$		480

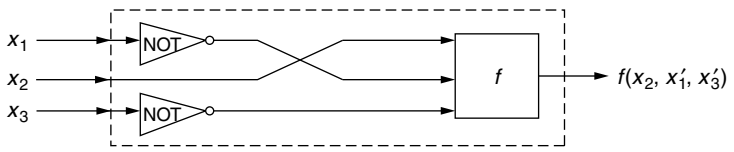


Figure 6.8. Permutation and complementation of inputs.

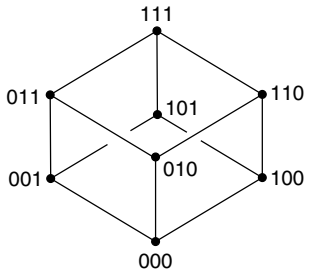


Figure 6.9. Cube of input values.

which sends proper rotations to 1 and improper rotations to -1; the kernel of ψ is the group of proper rotations of the cube which, by the morphism theorem, must be a normal subgroup of index 2. Therefore, the order of G is $2 \cdot 24 = 48$.

Example 6.7. Find the number of switching circuits involving three switches that are nonequivalent under permutation and complementation of the variables.

Solution. Each boolean function in three variables defines a coloring of the vertices of the cube of input values. A vertex is colored black if the function is 1 for the corresponding input value. It is colored white if the function takes the value 0 at that input value.

We can represent the complete symmetry group, G, of the cube by means of permutations of the vertices labeled 0, 1, 2, 3, 4, 5, 6, 7 in Figure 6.10. Since the group of proper rotations of the cube is a normal subgroup of index 2 in G, every element of G can be written as a proper rotation π or as $\pi \circ \rho$, where ρ is the reflection of the cube in its center.

There are 2^8 different switching functions of three variables, and Table 6.7 describes the number of circuits that are fixed by the action of each element of the group G on the eight inputs. For example, consider the element $g = (01) \circ (67) \circ (34) \circ (25)$. If a switching function f is fixed under the action of g, then the images of the input values corresponding to the vertices 0 and 1 must be the same; that is, $f(0, 0, 0) = f(0, 0, 1)$. Similarly, the images of the input values corresponding to the vertices 6 and 7 are the same, and $f(1, 1, 0) = f(1, 1, 1)$.

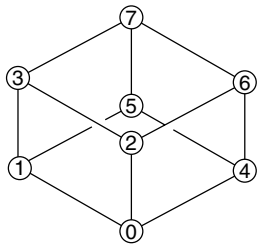


Figure 6.10. Labeling of the cube.

TABLE 6.7. Symmetries of a Cube Acting on the Three-Variable Switching Functions

Type of Element, g, in the Symmetry Group of the Cube	Order of g	Number, s, of Such Elements	$\lvert\text{Fix } g\rvert$	$s \cdot \lvert\text{Fix } g\rvert$
Proper rotations				
Identity	1	1	2^8	256
Rotation about a line joining midpoints of opposite edges [e.g., $(01)\circ(67)\circ(34)\circ(25)$]	2	6	2^4	96
Rotation about a line joining opposite vertices [e.g., $(124)\circ(365)\circ(0)\circ(7)$]	3	8	2^4	128
Rotation about a line joining centers of opposite faces [e.g., $(0264)\circ(1375)$]	4	6	2^2	24
Rotation about a line joining centers of opposite faces [e.g., $(06)\circ(24)\circ(17)\circ(35)$]	2	3	2^4	48
Improper rotations				
Reflection in the center [$\rho = (07)\circ(16)\circ(25)\circ(34)$]	2	1	2^4	16
Reflection in a diagonal plane [e.g., $(01)\circ(67)\circ(34)\circ(25)\circ\rho = (06)\circ(17)\circ(2)\circ(3)\circ(4)\circ(5)$]	2	6	2^6	384
Reflection and rotation [e.g., $(124)\circ(365)\circ\rho = (07)\circ(154623)$]	6	8	2^2	32
Reflection and rotation [e.g., $(0264)\circ(1375)\circ\rho = (0563)\circ(1472)$]	4	6	2^2	24
Reflection in a central plane [e.g., $(06)\circ(24)\circ(17)\circ(35)\circ\rho = (01)\circ(23)\circ(45)\circ(67)$]	2	3	2^4	48
		$\lvert G\rvert = 48$		1056

Also, $f(0,1,1) = f(1,0,0)$ and $f(0,1,0) = f(1,0,1)$. Hence the values of $f(0,0,0)$, $f(1,1,0)$, $f(0,1,1)$, and $f(0,1,0)$ can be chosen arbitrarily in 2^4 ways, and $\lvert\text{Fix } g\rvert = 2^4$. In general, if the function f is fixed under g, the images of the input values, corresponding to the vertices in any one cycle of g, must be the same. Hence $\lvert\text{Fix } g\rvert$ is 2^r, where r is the number of disjoint cycles in the permutation representation of g.

It follows from Table 6.7 and the Burnside theorem that the number of non-equivalent circuits is $1056/\lvert G\rvert = 1056/48 = 22$. □

TABLE 6.8. Number of Types of Switching Functions

Number of Switches, n	1	2	3	4	5
Number of boolean functions, 2^{2^n}	4	16	256	65,536	4,294,967,296
Nonequivalent functions under permutations of inputs	4	12	80	3,984	37,333,248
Nonequivalent functions under permutation and complementation of inputs	3	6	22	402	1,228,158
Nonequivalent functions under permutation and complementation of inputs and outputs	2	4	14	222	616,126

We can reduce this number slightly more by complementing the function as well as the variables; this corresponds to adding a NOT gate to the output. The group acting is now a combination of a cyclic group of order 2 with the complete symmetry groups of the cube.

The numbers of nonequivalent circuits for five or fewer switches given in Table 6.8 can be computed as in Example 6.7.

In 1951, the Harvard Computing Laboratory laboriously calculated all the nonequivalent circuits using four switches and the best way to design each of them. It was not until later that it was realized that the Pólya theory could be applied to this problem.

In many examples, it is quite difficult to calculate |Fix g| for every element g of the group G. Pólya's most important contribution to this theory of enumeration was to show how |Fix g| can be calculated, using what are called *cycle index polynomials*. This saves much individual calculation, and the results on nonequivalent boolean functions in Table 6.8 can easily be calculated. However, it is still a valuable exercise to tackle a few enumeration problems without using cycle index polynomials, since this gives a better understanding of the Pólya theory. For example, we see in Tables 6.3, 6.4, and 6.7 that |Fix g| is always of the form n^r, where r is the number of disjoint cycles in g.

Further information on the Pólya theory can be obtained from Biggs [15], Lidl and Pilz [10], or Stone [22].

EXERCISES

Find the number of different types of circular necklaces that could be made from the sets of beads described in Exercises 6.1 to 6.4, assuming that all the beads are used on one necklace.

6.1. Three black and three white beads.

6.2. Four black, three white, and one red bead.

6.3. Seven black and five white beads.

6.4. Five black, six white, and three red beads.

6.5. How many different circular necklaces containing ten beads can be made using beads of at most two colors?

6.6. Five neutral members and two members from each of two warring factions are to be seated around a circular armistice table. In how many nonequivalent ways, under the action of D_9, can they be seated if no two members of opposing factions sit next to each other?

6.7. How many different chemical compounds can be made by attaching H, CH_3, C_2H_5, or Cl radicals to the four bonds of a carbon atom? The radicals lie at the vertices of a regular tetrahedron, and the group is the tetrahedral group A_4.

6.8. How many different chemical compounds can be made by attaching H, CH_3, or OH radicals to each of the carbon atoms in the benzene ring of Figure 6.3? (Assume that all of the C–C bonds in the ring are equivalent.)

6.9. How many ways can the vertices of a cube be colored using, at most, three colors?

6.10. How many ways can the vertices of a regular tetrahedron be colored using, at most, n colors?

6.11. How many different tetrahedra can be made from n types of resistors when each edge contains one resistor?

6.12. How many ways can the faces of a regular dodecahedron be colored using, at most, n colors?

Find the number of different colorings of the faces of the solids described in Exercises 6.13 to 6.16.

6.13. A regular tetrahedron with two white faces and two black faces.

6.14. A cube with two white, one black, and three red faces.

6.15. A regular icosahedron with four black faces and 16 white faces.

6.16. A regular dodecahedron with five black faces, two white faces, and five green faces.

6.17. How many ways can the faces of a cube be colored with six different colors, if all the faces are to be a different color?

6.18. **(a)** Find the number of binary relations, on a set with four elements, that are not equivalent under permutations of the four elements.

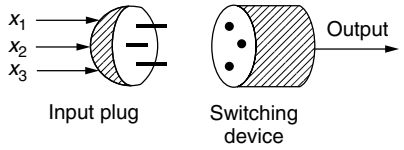


Figure 6.11

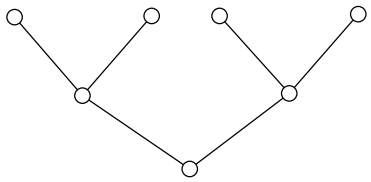


Figure 6.12

(b) Find the number of equivalence relations, on a set with four elements, that are not equivalent under permutations of the four elements.

6.19. How many different patchwork quilts, four patches long and three patches wide, can be made from five red and seven blue squares, assuming that the quilts cannot be turned over?

6.20. If the quilts in Exercise 6.19 could be turned over, how many different patterns are possible?

6.21. Find the number of ways of distributing three blue balls, two red balls, and four green balls into three piles.

6.22. If the cyclic group C_n, generated by g, operates on a set S, show that the number of orbits is

$$\frac{1}{n}\sum_{d/n} |\text{Fix } g^{n/d}| \cdot \phi(d),$$

where the Euler ϕ-function, $\phi(d)$, is the number of integers from 1 to d that are relatively prime to d. (See Exercises 4.55 and 4.56.)

6.23. Some transistor switching devices are sealed in a can with three input sockets at the vertices of an equilateral triangle. The three input wires are connected to a plug that will fit into the input sockets as shown in Figure 6.11. How many different cans are needed to produce any boolean function of three input variables?

6.24. How many different ways can the elements of the poset in Figure 6.12 be colored using, at most, n colors?

6.25. Verify that the number of nonequivalent switching functions of four variables, under permutation of the inputs, is 3984.

7

MONOIDS AND MACHINES

For many purposes, a group is too restrictive an algebraic concept, and we need a more general object. In the theory of machines, or automata theory, and in the mathematical study of languages and programming, algebraic objects arise naturally that have a single binary operation that is associative and has an identity. These are called *monoids*. The instructions to a digital machine consist of a sequence of input symbols that is fed into the machine. Two such sequences can be combined by following one by the other and, since this operation is associative, these input sequences form a monoid; the identity is the empty sequence that leaves the machine alone. Even though inverses do not necessarily exist in monoids, many of the general notions from group theory can be applied to these objects; for example, we can define subobjects, morphisms, and quotient objects.

MONOIDS AND SEMIGROUPS

A **monoid** $(M, \star)$ consists of a set M together with a binary operation $\star$ on M such that

(i) $a \star (b \star c) = (a \star b) \star c$ for all $a, b, c \in M$. (associativity)
(ii) There exists an **identity** $e \in M$ such that $a \star e = e \star a = a$ for all $a \in M$.

All groups are monoids. However, more general objects such as $(\mathbb{N}, +)$ and $(\mathbb{N}, \cdot)$, which do not have inverses, are also monoids.

A monoid $(M, \star)$ is called *commutative* if the operation $\star$ is commutative. The algebraic objects $(\mathbb{N}, +)$, $(\mathbb{N}, \cdot)$, $(\mathbb{Z}, +)$, $(\mathbb{Z}, \cdot)$, $(\mathbb{Q}, +)$, $(\mathbb{Q}, \cdot)$, $(\mathbb{R}, +)$, $(\mathbb{R}, \cdot)$, $(\mathbb{C}, +)$, $(\mathbb{C}, \cdot)$, $(\mathbb{Z}_n, +)$, and $(\mathbb{Z}_n, \cdot)$ are all commutative monoids.

However, $(\mathbb{Z}, -)$ is *not* a monoid because subtraction is not associative. In general, $(a - b) - c \neq a - (b - c)$.

Sometimes an algebraic object would be a monoid but for the fact that it lacks an identity element; such an object is called a *semigroup*. Hence a **semigroup**

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

$(S, \star)$ is just a set S together with an associative binary operation, $\star$. For example, $(\mathbb{P}, +)$ is a semigroup, but not a monoid, because the set of positive integers, $\mathbb{P}$, does not contain zero.

Just as one of the basic examples of a group consists of the permutations of any set, a basic example of a monoid is the set of transformations of any set. A transformation is just a function (not necessarily a bijection) from a set to itself. In fact, the analogue of Cayley's theorem holds for monoids, and it can be shown that every monoid can be represented as a transformation monoid.

Proposition 7.1. Let X be any set and let $X^X = \{f: X \to X\}$ be the set of all functions from X to itself. Then $(X^X, \circ)$ is a monoid, called the **transformation monoid** of X.

Proof. If $f, g \in X^X$, then the composition $f \circ g \in X^X$. Composition of functions is always associative, because if $f, g, h \in X^X$, then

$$(f \circ (g \circ h))(x) = f(g(h(x))) \quad \text{and} \quad ((f \circ g) \circ h)(x) = f(g(h(x)))$$

for all $x \in X$. The identity function $1_X: X \to X$ defined by $1_X(x) = x$ is the identity for composition. Hence $(X^X, \circ)$ is a monoid. □

Example 7.2. If $X = \{0, 1\}$, write out the table for the transformation monoid $(X^X, \circ)$.

Solution. X^X has four elements, e, f, g, h, defined as follows.

$$\begin{array}{llll} e(0) = 0 & f(0) = 0 & g(0) = 1 & h(0) = 1 \\ e(1) = 1 & f(1) = 0 & g(1) = 0 & h(1) = 1 \end{array}$$

The table for $(X^X, \circ)$ is shown in Table 7.1. For example, $g \circ f(0) = g(f(0)) = g(0) = 1$, and $g \circ f(1) = g(f(1)) = g(0) = 1$. Therefore, $g \circ f = h$. The other compositions can be calculated in a similar manner. □

Example 7.3. Prove that $(\mathbb{Z}, \star)$ is a commutative monoid, where $x \star y = 6 - 2x - 2y + xy$ for $x, y \in \mathbb{Z}$.

TABLE 7.1. Transformation Monoid of {0, 1}

$\circ$	e	f	g	h
e	e	f	g	h
f	f	f	f	f
g	g	h	e	f
h	h	h	h	h

Solution. For any $x, y \in \mathbb{Z}$, $x \star y \in \mathbb{Z}$, and $x \star y = y \star x$, so that $\star$ is a commutative binary operation on $\mathbb{Z}$. Now

$$\begin{aligned} x \star (y \star z) &= x \star (6 - 2y - 2z + yz) = 6 - 2x + (-2 + x)(6 - 2y - 2z + yz) \\ &= -6 + 4x + 4y + 4z - 2xy - 2xz - 2yz + xyz. \end{aligned}$$

Also,

$$\begin{aligned} (x \star y) \star z &= (6 - 2x - 2y + xy) \star z = 6 + (-2 + z)(6 - 2x - 2y + xy) - 2z \\ &= -6 + 4x + 4y + 4z - 2xy - 2xz - 2yz + xyz \\ &= x \star (y \star z). \end{aligned}$$

Hence $\star$ is associative.

Suppose that $e \star x = x$. Then $6 - 2e - 2x + ex = x$, and $6 - 2e - 3x + ex = 0$. This implies that $(x - 2)(e - 3) = 0$. Hence $e \star x = x$ for all $x \in \mathbb{Z}$ if and only if $e = 3$. Therefore, $(\mathbb{Z}, \star)$ is a commutative monoid with 3 as the identity. □

Since the operation in a monoid, $(M, \star)$, is associative, we can omit the parentheses when writing down a string of symbols combined by $\star$. We write the element $x_1 \star (x_2 \star x_3) = (x_1 \star x_2) \star x_3$ simply as $x_1 \star x_2 \star x_3$.

In any monoid $(M, \star)$ with identity e, the powers of any element $a \in M$ are defined by

$$a^0 = e, \quad a^1 = a, \quad a^2 = a \star a, \quad \ldots, \quad a^n = a \star a^{n-1} \quad \text{for } n \in \mathbb{N}.$$

The monoid $(M, \star)$ is said to be **generated** by the subset A if every element of M can be written as a finite combination of the powers of elements of A. That is, each element $m \in M$ can be written as

$$m = a_1^{r_1} \star a_2^{r_2} \star \cdots \star a_n^{r_n} \quad \text{for some } a_1, a_2, \ldots, a_n \in A.$$

For example, the monoid $(\mathbb{P}, \cdot)$ is generated by all the prime numbers. The monoid $(\mathbb{N}, +)$ is generated by the single element 1, since each element can be written as the sum of n copies of 1, where $n \in \mathbb{N}$. A monoid generated by one element is called a **cyclic monoid**.

A finite cyclic *group* is also a cyclic monoid. However, the infinite cyclic group $(\mathbb{Z}, +)$ is not a cyclic monoid; it needs at least two elements to generate it, for example, 1 and -1. Not all finite cyclic monoids are groups. For example, extending the notation of Chapter 3, let $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 1 & 4 & 3 \end{pmatrix} \in X^X$ where $X = \{1, 2, 3, 4\}$. Then $M = \{\varepsilon, \sigma, \sigma^2, \sigma^3\}$ is a cyclic monoid that is not a group because $\sigma^4 = \sigma^2$. More generally, the points in Figure 7.1 correspond to the elements of a cyclic monoid, and the arrows correspond to multiplication by the element c.

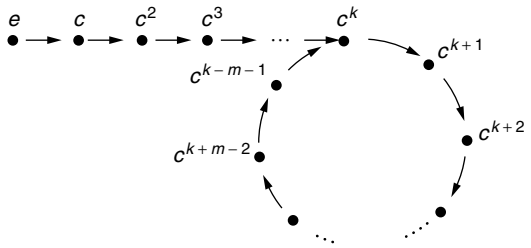


Figure 7.1. Finite cyclic monoid.

A computer receives its information from an input terminal that feeds in a sequence of symbols, usually binary digits consisting of 0's and 1's. If one sequence is fed in after another, the computer receives one long sequence that is the **concatenation** (or juxtaposition) of the two sequences. These input sequences together with the binary operation of concatenation form a monoid that is called the *free monoid generated by the input symbols*.

Let A be any set (sometimes called the **alphabet**), and let A^n be the set of n-tuples of elements in A. In this chapter, we write an n-tuple as a string of elements of A without any symbols between them. The elements of A^n are called **words of length** n from A. A word of length 0 is an empty string; this empty word is denoted by Λ. For example, if $A = \{a, b\}$, then $baabbaba \in A^8$, $A^0 = \{\Lambda\}$, and

$$A^3 = \{aaa, aab, aba, abb, baa, bab, bba, bbb\}.$$

Let $\mathrm{FM}(A)$ denote the set of all words from A, more formally

$$\mathrm{FM}(A) = A^0 \cup A \cup A^2 \cup A^3 \cup \cdots = \bigcup_{n=0}^{\infty} A^n.$$

Then $(\mathrm{FM}(A), \star)$ is called the **free monoid generated by** A, where the operation $\star$ is concatenation, and the identity is the empty word Λ. Another common notation for $\mathrm{FM}(A)$ is A^*.

If we do not include the empty word, Λ, we obtain the **free semigroup** generated by A; this is often denoted by A^+.

If α and β are words of length m and n, then $\alpha \star \beta$ is the word of length $m + n$ obtained by placing α to the left of β.

If A consists of a single element, a, then the monoid $\mathrm{FM}(A) = \{\Lambda, a, aa, aaa, aaaa, \ldots\}$ and, for example, $aaa \star aa = aaaaa$. This free monoid, generated by one element, is commutative.

If $A = \{0, 1\}$, then $\mathrm{FM}(A)$ consists of all the finite sequences of 0's and 1's,

$$\mathrm{FM}(A) = \{\Lambda, 0, 1, 00, 01, 10, 11, 000, 001, \ldots\}.$$

We have $010 \star 1110 = 0101110$ and $1110 \star 010 = 1110010$, so $\mathrm{FM}(A)$ is not commutative.

If $A = \{a, b, c, d, \ldots, y, z, \square, .\}$, the letters of the alphabet together with a space, $\square$, and a period, then

$$\text{the}\square\text{sky} \in \text{FM}(A) \quad \text{and} \quad \text{the}\square\text{sky} \star \text{is}\square\text{b} \star \text{lue.} = \text{the}\square\text{sky is}\square\text{blue.}$$

Of course, any nonsense string of letters is also in FM(A); for example, $pqb.a\square..\square xxu \in \text{FM}(A)$.

There is an important theorem that characterizes free monoids in terms of monoid morphisms. If $(M, \star)$ and $(N, \cdot)$ are two monoids, with identities e_M and e_N, respectively, then the function $f: M \to N$ is a **monoid morphism** from $(M, \star)$ to $(N, \cdot)$ if

(i) $f(x \star y) = f(x) \cdot f(y)$ for all $x, y \in M$.
(ii) $f(e_M) = e_N$.

A **monoid isomorphism** is simply a bijective monoid morphism.

For example, $f: (\mathbb{N}, +) \to (\mathbb{P}, \cdot)$ defined by $f(n) = 2^n$ is a monoid morphism because

$$f(n + m) = 2^{n+m} = 2^n \cdot 2^m = f(n) \cdot f(m) \quad \text{for all } m, n \in \mathbb{N}.$$

However, $f: \mathbb{N} \to \mathbb{N}$ defined by $f(x) = x^2$ is *not* a monoid morphism from $(\mathbb{N}, +)$ to $(\mathbb{N}, +)$. We have $f(x + y) = (x + y)^2$, whereas $f(x) + f(y) = x^2 + y^2$. Hence $f(1 + 1) = 4$, whereas $f(1) + f(1) = 2$.

Theorem 7.4. Let $(\text{FM}(A), \star)$ be the free monoid generated by A and let $i: A \to \text{FM}(A)$ be the function that maps each element a of A into the corresponding word of length 1, so that $i(a) = a$.

Then if $l: A \to M$ is any *function* into the underlying set of any monoid $(M, \cdot)$, there is a unique monoid morphism $h: (\text{FM}(A), \star) \to (M, \cdot)$ such that $h \circ i = l$. This is illustrated in Figure 7.2.

Proof. If h satisfies $h \circ i = l$, then h must be defined on words of length 1 by $h(a) = l(a)$. Once a morphism has been defined on its generators, it is determined completely as follows. Let α be a word of length $n \geqslant 2$ in FM(A). Write α as $\beta \star c$, where β is of length $n - 1$ and c is of length 1. Then we have $h(\alpha) = h(\beta \star c) = h(\beta) \cdot h(c) = h(\beta) \cdot l(c)$. Hence h can be determined by using induction on the word length. In fact, if $\alpha = a_1 a_2 \cdots a_n$, where $a_i \in A$, then $h(\alpha) = l(a_1) \cdot l(a_2) \ldots l(a_n)$. Finally, let $h(\Lambda)$ be the identity of M. $\square$

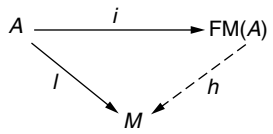


Figure 7.2. The function l factors through the free monoid FM(A).

FINITE-STATE MACHINES

We now look at mathematical models of sequential machines. These are machines that accept a finite set of inputs in sequential order. At any one time, the machine can be in one of a finite set of internal configurations or states. There may be a finite set of outputs. These outputs and internal states depend not only on the previous input but also on the stored information in the machine, that is, on the previous state of the machine. A pushbutton elevator is an example of such a machine. A digital computer is a very complex finite-state machine. It can be broken down into its component parts, each of which is also a machine. The *RS* and *JK* flip-flops, discussed in Exercises 2.69 and 2.70, are examples of two widely used components.

For simplicity, we only consider machines with a finite set of inputs and a finite set of states. We do not mention any outputs explicitly, because the state set can be enlarged, if necessary, to include any outputs. The states can be arranged so that a particular state always gives rise to a certain output.

A **finite-state machine**, (S, I, m) consists of a set of **states** $S = \{s_1, s_2, \ldots, s_n\}$, a set of **input values** $I = \{i_1, i_2, \ldots, i_t\}$, and a **transition function**

$$m: I \times S \to S,$$

which describes how each input value changes the states. If the machine is in state s_p and an input i_q is applied, the machine will change to state $m(i_q, s_p)$.

For example, consider a pushbutton elevator that travels between two levels, 1 and 2, and stops at the lower level 1 when not in use. We take the time for the elevator to travel from one level to the other to be the basic time interval, and the controlling machine can change states at the end of each interval. We allow the machine three inputs, so that $I = \{0, 1, 2\}$.

$$\text{input} = \begin{cases} 0 & \text{if no button is pressed in the preceding time interval} \\ 1 & \text{if button 1 only is pressed in the preceding time interval} \\ 2 & \text{if button 2 or both buttons are pressed} \\ & \quad \text{in the preceding time interval.} \end{cases}$$

Since the elevator is to stop at the bottom when not in use, we only consider states that end with the elevator going down. Let the set of states be

$$S = \{\text{stop, down, up–down, down–up–down}\}.$$

For example, in the "up–down" state, the elevator is traveling up, but must remember to come down. If no button is pressed or just button 1 is pressed while it is going up, the machine will revert to the "down" state when the elevator reaches level 2. On the other hand, if someone arrives at level 1 and presses button 2, the machine will change to the "down–up–down" state when the elevator reaches level 2.

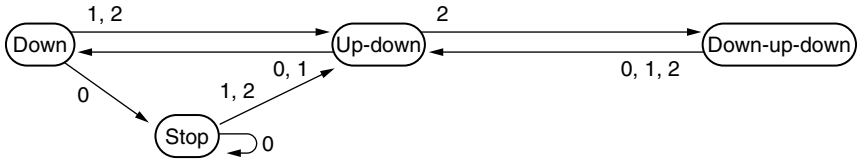


Figure 7.3. State diagram of the elevator.

The machine can be pictured by the **state diagram** in Figure 7.3. If the input i causes the machine to change from state s_p to state s_q, we draw an arrow labeled i from s_p to s_q in the diagram.

As another example, consider the following machine that checks the parity of the number of 1's fed into it. The set of states is $S = \{\text{start, even, odd}\}$, and the set of input values is $I = \{0, 1\}$. The function $m: I \times S \to S$ is described by Table 7.2, and the state diagram is given in Figure 7.4. If any sequence of 0's and 1's is fed into this machine, it will be in the even state if there is an even number of 1s in the sequence, and in an odd state otherwise.

Let I be the set of input values for any finite-state machine with state set S and function $m: I \times S \to S$. Each input value defines a function from the set of states to itself, the image of any state being the subsequent state produced by the given input. Hence we have a function

$$\tilde{m}: I \to S^S,$$

where S^S is the set of functions from S to itself, and $\tilde{m}(i): S \to S$ is defined by $[\tilde{m}(i)](s) = m(i, s)$.

TABLE 7.2. Transition Function of the Parity Checker

Initial State	Next State: Input 0	Next State: Input 1
Start	Even	Odd
Even	Even	Odd
Odd	Odd	Even

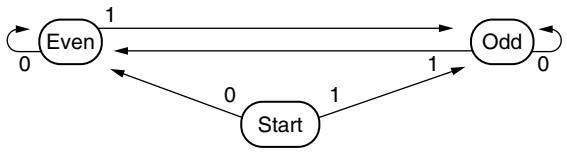


Figure 7.4. State diagram of the parity checker.

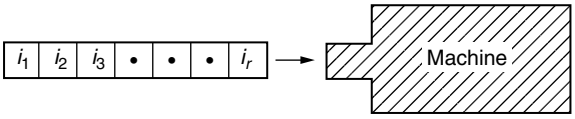


Figure 7.5. Input sequence being fed into a machine.

Any set of input values can be fed into the machine in sequence. The set of all such input sequences is the underlying set of the free monoid of input values, FM(I). By Theorem 7.4, the function $\tilde{m}: I \to S^S$ can be extended to a monoid morphism

$$h: (\mathrm{FM}(I), \star) \to (S^S, \circ),$$

where $h(i_1 i_2 \ldots i_r) = \tilde{m}(i_1) \circ \tilde{m}(i_2) \circ \cdots \circ \tilde{m}(i_r)$. Note that the input value i_r is fed into the machine first, and we can visualize this feeding of the input sequence in Figure 7.5. (The reader should be aware that many authors use the opposite convention in which the left input is fed into the machine first.)

For example, in the machine that checks the parity of the number of 1s in a sequence, the state set is $S = \{\text{start, even, odd}\}$ with functions

$$\tilde{m}: \{0, 1\} \to S^S \quad \text{and} \quad h: FM(\{0, 1\}) \to S^S.$$

The morphism h is defined by

$$h(\text{sequence}) = \begin{cases} \tilde{m}(0) & \text{if the sequence contains an even number of 1's} \\ \tilde{m}(1) & \text{if the sequence contains an odd number of 1's} \\ \text{identity function on } S & \text{if the sequence is empty.} \end{cases}$$

QUOTIENT MONOIDS AND THE MONOID OF A MACHINE

We have seen that different input sequences may have the same effect on a machine. For example, in the machine that checks the parity of the number of 1's in a sequence,

$$h(0101101) = h(0000) = h(11) = h(0);$$

thus the sequences 0101101, 0000, 11, and 0 cannot be distinguished by the machine.

In any machine with n states, the input sequences can have at most $|S^S| = n^n$ different effects. Since there are an infinite number of sequences in FM(I), there must always be many different input sequences that have the same effect.

The effect that an input has on a finite-state machine defines an equivalence relation on the input monoid FM(I). The monoid of a machine will be the quotient

monoid of FM(I) by this relation. It will always be a finite monoid with, at most, n^n elements. We first define the notion of a quotient monoid.

Suppose that R is an equivalence relation on a monoid $(M, \star)$. Then R is called a **congruence relation** on M if aRb implies that $(a \star c)R(b \star c)$ and $(c \star a)R(c \star b)$ for all $c \in M$. The **congruence class** containing the element $a \in M$ is the set

$$[a] = \{x \in M | xRa\}.$$

Proposition 7.5. If R is a congruence relation on the monoid $(M, \star)$, the quotient set $M/R = \{[a] | a \in M\}$ is a monoid under the operation defined by

$$[a] \star [b] = [a \star b].$$

This monoid is called the **quotient monoid** of M by R.

Proof. We first have to verify that the operation is well defined on congruence classes. Suppose that $[a] = [a']$ and $[b] = [b']$ so that aRa' and bRb'. Then $(a \star b)R(a \star b')$ and $(a \star b')R(a' \star b')$. Since R is transitive, $(a \star b)R(a' \star b')$ so $[a \star b] = [a' \star b']$. This shows that $\star$ is well defined on M/R. The associativity of $\star$ in M/R follows from the associativity of $\star$ in M. If e is the identity of M, then $[e]$ is the identity of M/R. Hence $(M/R, \star)$ is a monoid. □

Let (S, I, m) be a finite-state machine and let the effect of an input sequence be given by

$$h: \text{FM}(I) \to S^S.$$

Define the relation R on FM(I) by

$$\alpha R \beta \quad \text{if and only if} \quad h(\alpha) = h(\beta).$$

This is easily verified to be an equivalence relation. Furthermore, it is a congruence relation on the free monoid (FM(I), $\star$), because if $\alpha R \beta$, then $h(\alpha) = h(\beta)$, and $h(\alpha \star \gamma) = h(\alpha) \circ h(\gamma) = h(\beta) \circ h(\gamma) = h(\beta \star \gamma)$; thus $(\alpha \star \gamma)R(\beta \star \gamma)$, and similarly, $(\gamma \star \alpha)R(\gamma \star \beta)$.

The quotient monoid (FM(I)$/R, \star$) is called the **monoid of the machine** (S, I, m).

We can apply the same construction to the free semigroup of input sequences to obtain the **semigroup of the machine**.

The monoid of a machine reflects the capability of the machine to respond to the input sequences. There are an infinite number of sequences in FM(I), whereas the number of elements in the quotient monoid is less than or equal to n^n. Two sequences are in the same congruence class if and only if they have the same effect on the machine.

A morphism theorem for monoids can be proved in a similar way to the morphism theorem for groups (Theorem 4.25; see Exercise 7.24). Applying this

to the monoid morphism $h: \mathrm{FM}(I) \to S^S$, it follows that the quotient monoid $\mathrm{FM}(I)/R$ is isomorphic to $\mathrm{Im}\, h$. This isomorphism assigns to each congruence class a unique transition between states.

Example 7.6. Draw the state diagram and find the monoid of the following machine (S, I, m). The machine has two states, s_0 and s_1, and two input symbols, 0 and 1. The effects of the input symbols are given by the functions $h(0), h(1): S \to S$, defined in Table 7.3.

Solution. Let us calculate the effect of inputs of length 2. We have $h(ij) = h(i) \circ h(j)$, where j is fed into the machine first. It follows from Tables 7.3 and 7.4 that $h(00) = h(01) = h(0)$ and $[00] = [01] = [0]$ in the monoid of the machine. There are only four functions from $\{s_0, s_1\}$ to $\{s_0, s_1\}$, and these are $h(0), h(1), h(10)$, and $h(11)$. Hence the monoid of the machine consists of the four congruence classes [0], [1], [10], and [11]. The table of this quotient monoid is given in Table 7.5, and the state diagram is given in Figure 7.6. For example,

TABLE 7.3

Initial State	Next State	
	$h(0)$	$h(1)$
s_0	s_0	s_1
s_1	s_0	s_0

TABLE 7.4

Initial State	End State			
	$h(00)$	$h(01)$	$h(10)$	$h(11)$
s_0	s_0	s_0	s_1	s_0
s_1	s_0	s_0	s_1	s_1

TABLE 7.5. Monoid of the Machine

$\star$	[0]	[1]	[10]	[11]
[0]	[0]	[0]	[0]	[0]
[1]	[10]	[11]	[0]	[1]
[10]	[10]	[10]	[10]	[10]
[11]	[0]	[1]	[10]	[11]

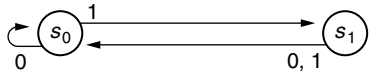


Figure 7.6. State diagram.

$[1] \star [10] = [110]$. Since $h(110)(s_0) = s_0$ and $h(110)(s_1) = s_0$, it follows that $[110] = [0]$. Notice that [11] is the identity; thus, in the monoid of the machine, $[\Lambda] = [11]$. □

Example 7.7. Describe the monoid of the machine $(\{\text{start, even, odd}\}, \{0, 1\}, m)$ that determines the parity of the number of 1's in the input.

Solution. We have already seen that any input sequence with an even number of 1's has the same effect as 0 and that any sequence with an odd number of 1's has the same effect as 1. It follows from Table 7.6 that the monoid of the machine contains the three elements $[\Lambda]$, [0], and [1]. The table for this monoid is given in Table 7.7. □

Finite-state machines can easily be designed to recognize certain types of input sequences. For example, most numbers inside a computer are in binary form and have a check digit attached to them so that there is always an even number of 1's in each sequence. This is used to detect any machine errors (see Chapter 14). A finite-state machine like Example 7.7 can be used to perform a parity check on all the sequences of numbers in the computer. The machine can be designed to signal a parity check error whenever it ends in the "odd" state.

Let us now look at a machine that will recognize the pattern 010 in any binary input sequence that is fed into the machine. Figure 7.7 is the state diagram of such a machine. If the machine is initiated in state s_1, it will be in state s_4 if and only if the preceding inputs were 010, and in this case, the machine sends an output signal.

This machine has four states; thus the total possible number of different functions between states is $4^4 = 256$. Table 7.8 shows that the input sequences of length 0, 1, and 2 all have different effects on the various states. However, seven of the eight sequences of length 3 have the same effect as sequences of length

TABLE 7.6

Initial State	Next State		
	$h(\Lambda)$	$h(0)$	$h(1)$
Start	Start	Even	Odd
Even	Even	Even	Odd
Odd	Odd	Odd	Even

TABLE 7.7. Monoid of the Parity Checker Machine

$\star$	$[\Lambda]$	[0]	[1]
$[\Lambda]$	$[\Lambda]$	[0]	[1]
[0]	[0]	[0]	[1]
[1]	[1]	[1]	[0]

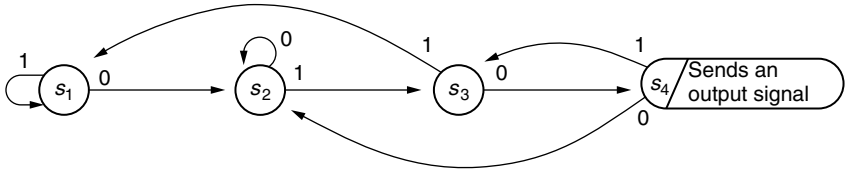


Figure 7.7. State diagram of a machine that recognizes the sequence 010.

TABLE 7.8. Effects of the Input Sequences on the States of the Machine

Initial State	End State																	
	Λ	0	1	00	01	10	11	000	001	010	011	100	101	110	111	0010	1010
s_1	s_1	s_2	s_1	s_2	s_2	s_3	s_1	s_2	s_2	s_4	s_2	s_3	s_3	s_1	s_1	s_2	s_3
s_2	s_2	s_2	s_3	s_2	s_4	s_3	s_1	s_2	s_2	s_4	s_2	s_3	s_3	s_1	s_1	s_2	s_3
s_3	s_3	s_4	s_1	s_2	s_2	s_3	s_1	s_2	s_2	s_4	s_2	s_3	s_3	s_1	s_1	s_2	s_3
s_4	s_4	s_2	s_3	s_2	s_4	s_3	s_1	s_2	s_2	s_4	s_2	s_3	s_3	s_1	s_1	s_2	s_3

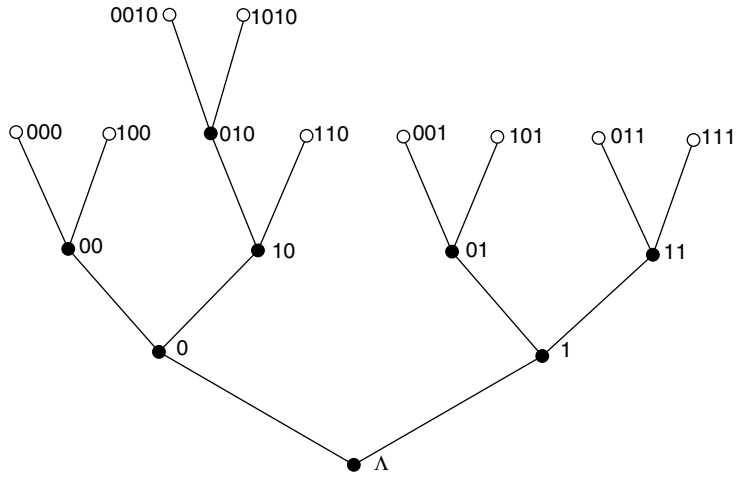


Figure 7.8. Tree diagram of input sequences.

2. The only input sequence with a different effect is 010, the sequence that the machine is designed to recognize. Therefore, the only sequences of length 4 that we check are those whose initial inputs are 010, namely, 0010 and 1010.

We can use the *tree diagram* in Figure 7.8 to check that we have covered all the possible transition functions obtainable by any input sequence. We label the nodes of the tree by input sequences. At any node α, there will be two upward branches ending in the nodes $0 \star \alpha$ and $1 \star \alpha$, corresponding to the two input symbols. We prune the tree at node α, if α gives rise to the same transition function as another

TABLE 7.9. Monoid of the Machine That Recognizes 010

$\star$	[Λ]	[0]	[1]	[00]	[01]	[10]	[11]	[010]
[Λ]	[Λ]	[0]	[1]	[00]	[01]	[10]	[11]	[010]
[0]	[0]	[00]	[01]	[00]	[00]	[010]	[00]	[00]
[1]	[1]	[10]	[11]	[10]	[10]	[11]	[11]	[10]
[00]	[00]	[00]	[00]	[00]	[00]	[00]	[00]	[00]
[01]	[01]	[010]	[00]	[010]	[010]	[00]	[00]	[010]
[10]	[10]	[10]	[10]	[10]	[10]	[10]	[10]	[10]
[11]	[11]	[11]	[11]	[11]	[11]	[11]	[11]	[11]
[010]	[010]	[010]	[010]	[010]	[010]	[010]	[010]	[010]

node β in the tree. The tree must eventually stop growing because there are only a finite number of transition functions. Every input sequence has the same effect as one of the solid black nodes in Figure 7.8. These nodes provide a complete set of representatives for the monoid of the machine.

Therefore, the monoid of the machine that recognizes the sequence 010 contains only eight elements: [Λ], [0], [1], [00], [01], [10], [11], and [010], out of a possible 256 transition functions between states. Its table is given in Table 7.9.

For further reading on the mathematical structure of finite-state machines and automata see Hopcroft et al. [18], Kolman [20], or Stone [22].

EXERCISES

Are the structures described in Exercises 7.1 to 7.13 semigroups or monoids or neither? Give the identity of each monoid.

7.1. ($\mathbb{N}$, gcd).

7.2. ($\mathbb{Z}$, [), where $a[b = a$.

7.3. ($\mathbb{R}$, $\star$), where $x \star y = \sqrt{x^2 + y^2}$.

7.4. ($\mathbb{R}$, $\star$), where $x \star y = \sqrt[3]{x^3 + y^3}$.

7.5. ($\mathbb{Z}_3$, $-$).

7.6. ($\mathbb{R}$, | |), where | | is the absolute value.

7.7. ($\mathbb{Z}$, max), where max (m, n) is the larger of m and n.

7.8. ($\mathbb{Z}$, $\star$), where $x \star y = x + y + xy$.

7.9. (S, gcd), where $S = \{1, 2, 3, 4, 5, 6\}$.

7.10. (X, max), where X is the set of real-valued functions on the unit interval [0,1] and if $f, g \in X$, then max (f, g) is the function on X defined by

$$\max(f, g)(x) = \max(f(x), g(x)).$$

7.11. (T, lcm) where $T = \{1, 2, 4, 5, 10, 20\}$.

7.12. The set of all relations on a set X, where the composition of two relations R and S is the relation RS defined by $xRSz$ if and only if for some $y \in X$, xRy and ySz.

7.13. $(\{a, b, c\}, \star)$, where the table for $\star$ is given in Table 7.10.

TABLE 7.10

$\star$	a	b	c
a	a	b	c
b	b	a	a
c	c	a	a

Write out the tables for the monoids and semigroups described in Exercises 7.14 to 7.17.

7.14. (S, gcd), where $S = \{1, 2, 3, 4, 6, 8, 12, 24\}$.

7.15. (T, gcd), where $T = \{1, 2, 3, 4\}$.

7.16. $(X^X, \circ)$, where $X = \{1, 2, 3\}$.

7.17. $(\{e, c, c^2, c^3, c^4\}, \cdot)$, where multiplication by c is indicated by an arrow in Figure 7.9.

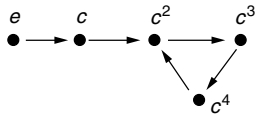


Figure 7.9

7.18. Find all the commutative monoids on the set $S = \{e, a, b\}$ with identity e.

7.19. Are all the elements of the free semigroup generated by {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} simply the nonnegative integers written in the base 10?

7.20. A **submonoid** of a monoid $(M, \cdot)$ is a subset N of M containing the identity and such that $x \cdot y \in N$, for all $x, y \in N$. Find all the submonoids of the monoid given in Exercise 7.17.

7.21. Prove that there is a monoid isomorphism between $(\text{FM}(\{a\}), \star)$ and $(\mathbb{N}, +)$.

7.22. **(Representation theorem for monoids)** Prove that any monoid $(M, \star)$ is isomorphic to a submonoid of $(M^M, \circ)$. This gives a representation of any monoid as a monoid of transformations.

7.23. Prove that any cyclic monoid is either isomorphic to $(\mathbb{N}, +)$ or is isomorphic to a monoid of the form shown in Figure 7.1, for some values of k and m.

7.24. **(Morphism theorem for monoids)** Let $f: (M, \star) \to (N, \cdot)$ be a morphism of monoids. Let R be the relation on M defined by $m_1 R m_2$ if and only if

$f(m_1) = f(m_2)$. Prove that the quotient monoid $(M/R, \star)$ is isomorphic to the submonoid $(\mathrm{Im} f, \cdot)$ of $(N, \cdot)$. (See Exercise 7.20.)

7.25. An **automorphism** of a monoid M is an isomorphism from M to itself. Prove that the set of all automorphisms of a monoid M forms a group under composition.

7.26. A machine has three states, s_1, s_2, and s_3 and two input symbols, α and β. The effect of the input symbols on the states is given by Table 7.11. Draw the state diagram and find the monoid of this machine.

TABLE 7.11

Initial State	Next State	
	$h(\alpha)$	$h(\beta)$
s_1	s_1	s_1
s_2	s_3	s_1
s_3	s_2	s_1

7.27. Prove that every finite monoid is the monoid of some finite-state machine.

For Exercises 7.28 to 7.30, draw state diagrams of machines with the given input set, I, that will recognize the given sequence.

7.28. 1101, where $I = \{0, 1\}$.

7.29. 0101, where $I = \{0, 1\}$.

7.30. 2131, where $I = \{1, 2, 3\}$.

Which of the relations described in Exercises 7.31 to 7.34 are congruence relations on the monoid $(\mathbb{N}, +)$*? Find the quotient monoid when the relation is a congruence relation.*

7.31. aRb if $a - b$ is even.

7.32. aRb if $a > b$.

7.33. aRb if $a = 2^r b$ for some $r \in \mathbb{Z}$.

7.34. aRb if $10|(a - b)$.

The machines in Tables 7.12, 7.13, and 7.14 have state set $S = \{s_1, s_2, s_3\}$ *and input set* $I = \{0, 1\}$.

7.35. Draw the table of the monoid of the machine defined by Table 7.12.

TABLE 7.12

Initial State	Next State	
	$h(0)$	$h(1)$
s_1	s_2	s_1
s_2	s_1	s_2
s_3	s_3	s_2

7.36. Draw the table of the monoid of the machine defined by Table 7.13.

TABLE 7.13

Initial State	Next State	
	$h(0)$	$h(1)$
s_1	s_2	s_1
s_2	s_3	s_1
s_3	s_3	s_2

7.37. Find the number of elements in the monoid of the machine defined by Table 7.14.

TABLE 7.14

Initial State	Next State	
	$h(0)$	$h(1)$
s_1	s_2	s_1
s_2	s_3	s_3
s_3	s_1	s_1

7.38. Find the number of elements in the semigroup of the machine, given by Figure 7.3, that controls the elevator.

7.39. Find the monoid of the machine in Figure 7.10.

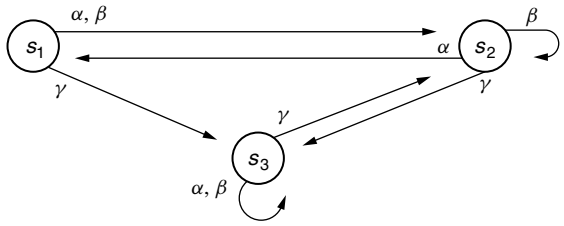


Figure 7.10

7.40. A **serial adder**, illustrated in Figure 7.11, is a machine that adds two numbers in binary form. The two numbers are fed in together, one digit at a time, starting from the right end. Their sum appears as the output. The machine has input symbols 00, 01, 10, and 11, corresponding to the rightmost digits of the numbers. Figure 7.12 gives the state diagram of such a machine, where the symbol "s_{ij}/j" indicates that the machine is in state s_{ij} and emits an output j. The carry digit is the number i of the state s_{ij}. Find the monoid of this machine.

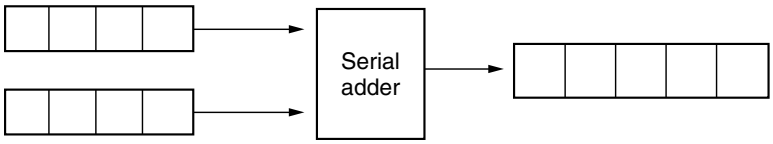


Figure 7.11

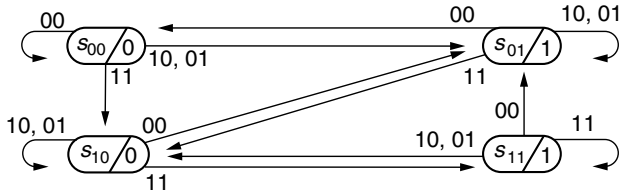
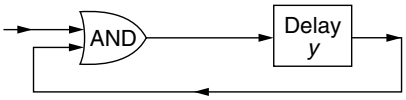


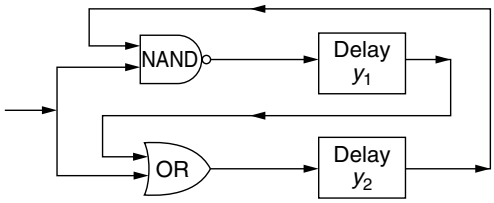
Figure 7.12. State diagram of the serial adder.

The circuits in Exercises 7.41 to 7.44 represent the internal structures of some finite-state machines constructed from transistor circuits. These circuits are controlled by a clock, and the rectangular boxes denote delays of one time unit. The input symbols are 0 and 1 and are fed in at unit time intervals. The internal states of the machines are described by the contents of the delays. Draw the state diagram and find the elements in the semigroup of each machine.

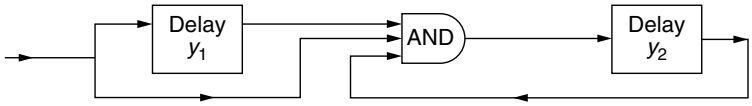
7.41.



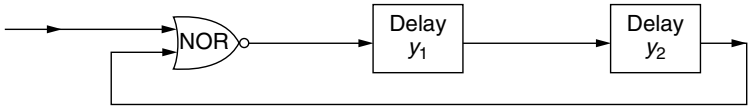
7.42.



7.43.



7.44.



7.45. In the spring, a plant bud has to have the right conditions in order to develop. One particular bud has to have a rainy day followed by two warm days, without being interrupted by cool or freezing days, in order to develop. Furthermore, if a freezing day occurs after the bud has developed, the bud dies. Draw a state diagram for such a bud using the input symbols R, W, C, F to stand for rainy, warm, cool, and freezing days, respectively. What is the number of elements in the resulting monoid of this bud?

7.46. A dog can either be passive, angry, frightened, or angry and frightened, in which case he bites. If you give him a bone, he becomes passive. If you remove one of his bones, he becomes angry, and, if he is already frightened, he will bite you. If you threaten him, he becomes frightened, but, if he is already angry, he will bite. Write out the table of the monoid of the dog.

8

RINGS AND FIELDS

The familiar number systems of the real or complex numbers contain two basic binary operations, addition and multiplication. Group theory is not sufficient to capture all of the algebraic structure of these number systems, because a group deals with only one binary operation. It is possible to consider the integers as a group $(\mathbb{Z}, +)$ and the nonzero integers as a monoid $(\mathbb{Z}^*, \cdot)$, but this still neglects the relation between addition and multiplication, namely, the fact that multiplication is distributive over addition. We therefore consider algebraic structures with two binary operations modeled after these number systems. A ring is a structure that has the minimal properties we would expect of addition and multiplication. A field is a more specialized ring in which division by nonzero elements is always possible.

In this chapter we look at the basic properties of rings and fields and consider many examples. In later chapters we construct new number systems with properties similar to the familiar systems.

RINGS

A **ring** $(R, +, \cdot)$ is a set R, together with two binary operations $+$ and $\cdot$ on R satisfying the following axioms. For any elements $a, b, c \in R$,

(i)	$(a + b) + c = a + (b + c)$.	(associativity of addition)
(ii)	$a + b = b + a$.	(commutativity of addition)
(iii)	there exists $0 \in R$, called the *zero*, such that $a + 0 = a$.	(existence of an additive identity)
(iv)	there exists $(-a) \in R$ such that $a + (-a) = 0$.	(existence of an additive inverse)
(v)	$(a \cdot b) \cdot c = a \cdot (b \cdot c)$.	(associativity of multiplication)
(vi)	there exists $1 \in R$ such that $1 \cdot a = a \cdot 1 = a$.	(existence of multiplicative identity)

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

(vii) $a \cdot (b + c) = a \cdot b + a \cdot c$ and (distributivity)
$(b + c) \cdot a = b \cdot a + c \cdot a$.

Axioms (i)–(iv) are equivalent to saying that $(R, +)$ is an abelian group, and axioms (v) and (vi) are equivalent to saying that $(R, \cdot)$ is a monoid.

The ring $(R, +, \cdot)$ is called a **commutative ring** if, in addition,

(viii) $a \cdot b = b \cdot a$ for all $a, b \in R$. (commutativity of multiplication)

The integers under addition and multiplication satisfy all of the axioms above, so that $(\mathbb{Z}, +, \cdot)$ is a commutative ring. Also, $(\mathbb{Q}, +, \cdot)$, $(\mathbb{R}, +, \cdot)$, and $(\mathbb{C}, +, \cdot)$ are all commutative rings. If there is no confusion about the operations, we write only R for the ring $(R, +, \cdot)$. Therefore, the rings above would be referred to as $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, or $\mathbb{C}$. Moreover, if we refer to a ring R without explicitly defining its operations, it can be assumed that they are addition and multiplication.

Many authors do not require a ring to have a multiplicative identity, and most of the results we prove can be verified to hold for these objects as well. Exercise 8.49 shows that such an object can always be embedded in a ring that does have a multiplicative identity.

The set of all $n \times n$ square matrices with real coefficients forms a ring $(M_n(\mathbb{R}), +, \cdot)$, which is *not commutative* if $n > 1$, because matrix multiplication is not commutative.

The elements "even" and "odd" form a commutative ring $(\{\text{even}, \text{odd}\}, +, \cdot)$ where the operations are given by Table 8.1. "Even" is the zero of this ring, and "odd" is the multiplicative identity. This is really a special case of the following example when $n = 2$.

Example 8.1. Show that $(\mathbb{Z}_n, +, \cdot)$ is a commutative ring, where addition and multiplication on congruence classes, modulo n, are defined by the equations $[x] + [y] = [x + y]$ and $[x] \cdot [y] = [xy]$.

Solution. It follows from Example 4.19, that $(\mathbb{Z}_n, +)$ is an abelian group.

Since multiplication on congruence classes is defined in terms of representatives, it must be verified that it is well defined. Suppose that $[x] = [x']$ and $[y] = [y']$, so that $x \equiv x'$ and $y \equiv y' \mod n$. This implies that $x = x' + kn$ and $y = y' + ln$ for some $k, l \in \mathbb{Z}$. Now $x \cdot y = (x' + kn) \cdot (y' + ln) = x' \cdot y' + (ky' + lx' + kln)n$, so $x \cdot y \equiv x' \cdot y' \mod n$ and hence $[x \cdot y] = [x' \cdot y']$. This shows that multiplication is well defined.

TABLE 8.1. Ring of Odd and Even Integers

$+$	Even	Odd
Even	even	odd
Odd	odd	even

$\cdot$	Even	Odd
Even	even	even
Odd	even	odd

The remaining axioms now follow from the definitions of addition and multiplication and from the properties of the integers. The zero is [0], and the unit is [1]. The left distributive law is true, for example, because

$$\begin{aligned}[x]\cdot([y]+[z]) &= [x]\cdot[y+z] = [x\cdot(y+z)]\\ &= [x\cdot y + x\cdot z] \qquad \text{by distributivity in } \mathbb{Z}\\ &= [x\cdot y]+[x\cdot z] = [x]\cdot[y]+[x]\cdot[z]. \qquad \square\end{aligned}$$

Example 8.2. Construct the addition and multiplication tables for the ring $(\mathbb{Z}_5, +, \cdot)$.

Solution. We denote the congruence class $[x]$ just by x. The tables are given in Table 8.2. $\square$

Example 8.3. Show that $(\mathbb{Q}(\sqrt{2}), +, \cdot)$ is a commutative ring where $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \in \mathbb{R} | a, b \in \mathbb{Q}\}$.

Solution. The set $\mathbb{Q}(\sqrt{2})$ is a subset of $\mathbb{R}$, and the addition and multiplication is the same as that of real numbers. First, we check that $+$ and $\cdot$ are binary operations on $\mathbb{Q}(\sqrt{2})$. If $a, b, c, d \in \mathbb{Q}$, we have

$$(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$$

since $(a + c)$ and $(b + d) \in \mathbb{Q}$. Also,

$$(a + b\sqrt{2}) \cdot (c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$$

since $(ac + 2bd)$ and $(ad + bc) \in \mathbb{Q}$.

We now check that axioms (i)–(viii) of a commutative ring are valid in $\mathbb{Q}(\sqrt{2})$.

(i) Addition of real numbers is associative.
(ii) Addition of real numbers is commutative.
(iii) The zero is $0 = 0 + 0\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.
(iv) The additive inverse of $a + b\sqrt{2}$ is $(-a) + (-b)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$, since $(-a)$ and $(-b) \in \mathbb{Q}$.

TABLE 8.2. Ring $(\mathbb{Z}_5, +, \cdot)$

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

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

(v) Multiplication of real numbers is associative.
(vi) The multiplicative identity is $1 = 1 + 0\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.
(vii) The distributive axioms hold for real numbers and hence hold for elements of $\mathbb{Q}(\sqrt{2})$.
(viii) Multiplication of real numbers is commutative. □

We have already investigated one algebraic system with two binary operations: a boolean algebra. The boolean algebra of subsets of a set is not a ring under the operations of union and intersection, because neither of these operations has inverses. However, the symmetric difference does have an inverse, and we can make a boolean algebra into a ring using this operation and the operation of intersection.

Example 8.4. $(\mathcal{P}(X), \Delta, \cap)$ is a commutative ring for any set X.

Solution. The axioms (i)–(viii) of a commutative ring follow from Propositions 2.1 and 2.3. The zero is $\emptyset$, and the identity is X. □

In the ring above, $A \cap A = A$ for every element A in the ring. Such rings are called **boolean rings**, since they are all derivable from boolean algebras (see Exercise 8.13).

Example 8.5. Construct the tables for the ring $(\mathcal{P}(X), \Delta, \cap)$, where $X = \{a, b, c\}$.

Solution. Let $A = \{a\}$, $B = \{b\}$, and $C = \{c\}$, so that $\overline{A} = \{b, c\}$, $\overline{B} = \{a, c\}$, and $\overline{C} = \{a, b\}$. Therefore, $\mathcal{P}(X) = \{\emptyset, A, B, C, \overline{A}, \overline{B}, \overline{C}, X\}$. The tables for the symmetric difference and intersection are given in Table 8.3. □

The following properties are useful in manipulating elements of any ring.

Proposition 8.6. If $(R, +, \cdot)$ is a ring, then for all $a, b \in R$:

(i) $a \cdot 0 = 0 \cdot a = 0$.
(ii) $a \cdot (-b) = (-a) \cdot b = -(a \cdot b)$.
(iii) $(-a) \cdot (-b) = a \cdot b$.
(iv) $(-1) \cdot a = -a$.
(v) $(-1) \cdot (-1) = 1$.

Proof. (i) By distributivity, $a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0$. Adding $-(a \cdot 0)$ to each side, we obtain $0 = a \cdot 0$. Similarly, $0 \cdot a = 0$.

(ii) Compute $a \cdot (-b) + a \cdot b = a \cdot (-b + b) = a \cdot 0 = 0$, using (i). Therefore, $a \cdot (-b) = -(a \cdot b)$. Similarly, $(-a) \cdot b = -(a \cdot b)$.

(iii) We have $(-a) \cdot (-b) = -(a \cdot (-b)) = -(-(a \cdot b)) = a \cdot b$ by (ii) and Proposition 3.7.

TABLE 8.3. Ring $\mathcal{P}(\{a,b,c\})$

Δ	$\emptyset$	A	B	C	$\overline{A}$	$\overline{B}$	$\overline{C}$	X
$\emptyset$	$\emptyset$	A	B	C	$\overline{A}$	$\overline{B}$	$\overline{C}$	X
A	A	$\emptyset$	$\overline{C}$	$\overline{B}$	X	C	B	$\overline{A}$
B	B	$\overline{C}$	$\emptyset$	$\overline{A}$	C	X	A	$\overline{B}$
C	C	$\overline{B}$	$\overline{A}$	$\emptyset$	B	A	X	$\overline{C}$
$\overline{A}$	$\overline{A}$	X	C	B	$\emptyset$	$\overline{C}$	$\overline{B}$	A
$\overline{B}$	$\overline{B}$	C	X	A	$\overline{C}$	$\emptyset$	$\overline{A}$	B
$\overline{C}$	$\overline{C}$	B	A	X	$\overline{B}$	$\overline{A}$	$\emptyset$	C
X	X	$\overline{A}$	$\overline{B}$	$\overline{C}$	A	B	C	$\emptyset$

$\cap$	$\emptyset$	A	B	C	$\overline{A}$	$\overline{B}$	$\overline{C}$	X
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
A	$\emptyset$	A	$\emptyset$	$\emptyset$	$\emptyset$	A	A	A
B	$\emptyset$	$\emptyset$	B	$\emptyset$	B	$\emptyset$	B	B
C	$\emptyset$	$\emptyset$	$\emptyset$	C	C	C	$\emptyset$	C
$\overline{A}$	$\emptyset$	$\emptyset$	B	C	$\overline{A}$	C	B	$\overline{A}$
$\overline{B}$	$\emptyset$	A	$\emptyset$	C	C	$\overline{B}$	A	$\overline{B}$
$\overline{C}$	$\emptyset$	A	B	$\emptyset$	B	A	$\overline{C}$	$\overline{C}$
X	$\emptyset$	A	B	C	$\overline{A}$	$\overline{B}$	$\overline{C}$	X

(iv) By (ii), $(-1)\cdot a = -(1\cdot a) = -a$.
(v) By (iii), $(-1)\cdot(-1) = 1\cdot 1 = 1$. □

Proposition 8.7. If $0 = 1$, the ring contains only one element and is called the **trivial ring**. All other rings are called **nontrivial**.

Proof. For any element, a, in a ring in which $0 = 1$, we have $a = a\cdot 1 = a\cdot 0 = 0$. Therefore, the ring contains only the element 0. It can be verified that this forms a ring with the operations defined by $0 + 0 = 0$ and $0\cdot 0 = 0$. □

INTEGRAL DOMAINS AND FIELDS

One very useful property of the familiar number systems is the fact that if $ab = 0$, then either $a = 0$ or $b = 0$. This property allows us to cancel nonzero elements because if $ab = ac$ and $a \neq 0$, then $a(b - c) = 0$, so $b = c$. However, this property does not hold for all rings. For example, in $\mathbb{Z}_4$, we have $[2]\cdot[2] = [0]$, and we cannot always cancel since $[2]\cdot[1] = [2]\cdot[3]$, but $[1] \neq [3]$.

If $(R, +, \cdot)$ is a commutative ring, a nonzero element $a \in R$ is called a **zero divisor** if there exists a nonzero element $b \in R$ such that $a\cdot b = 0$. A nontrivial commutative ring is called an **integral domain** if it has no zero divisors. Hence

a nontrivial commutative ring is an integral domain if $a \cdot b = 0$ always implies that $a = 0$ or $b = 0$.

As the name implies, the integers form an integral domain. Also, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are integral domains. However, $\mathbb{Z}_4$ is *not*, because [2] is a zero divisor. Neither is $(\mathcal{P}(X), \Delta, \cap)$, because every nonempty proper subset of X is a zero divisor. $M_n(\mathbb{R})$ is *not* an integral domain (for example, $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}^2 = 0$).

Proposition 8.8. If a is a nonzero element of an integral domain R and $a \cdot b = a \cdot c$, then $b = c$.

Proof. If $a \cdot b = a \cdot c$, then $a \cdot (b - c) = a \cdot b - a \cdot c = 0$. Since R is an integral domain, it has no zero divisors. Since $a \neq 0$, it follows that $(b - c) = 0$. Hence $b = c$. □

Generally speaking, it is possible to add, subtract, and multiply elements in a ring, but it is not always possible to divide. Even in an integral domain, where elements can be canceled, it is not always possible to divide by nonzero elements. For example, if $x, y \in \mathbb{Z}$, then $2x = 2y$ implies that $x = y$, but not all elements in $\mathbb{Z}$ can be divided by 2.

The most useful number systems are those in which we can divide by nonzero elements. A **field** is a ring in which the nonzero elements form an abelian group under multiplication. In other words, a field is a nontrivial commutative ring R satisfying the following extra axiom.

(ix) For each nonzero element $a \in R$ there exists $a^{-1} \in R$ such that $a \cdot a^{-1} = 1$.

The rings $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ are all fields, but the integers do not form a field.

Proposition 8.9. Every field is an integral domain; that is, it has no zero divisors.

Proof. Let $a \cdot b = 0$ in a field F. If $a \neq 0$, there exists an inverse $a^{-1} \in F$ and $b = (a^{-1} \cdot a) \cdot b = a^{-1}(a \cdot b) = a^{-1} \cdot 0 = 0$. Hence either $a = 0$ or $b = 0$, and F is an integral domain. □

Theorem 8.10. A finite integral domain is a field.

Proof. Let $D = \{x_0, x_1, x_2, \ldots, x_n\}$ be a finite integral domain with x_0 as 0 and x_1 as 1. We have to show that every nonzero element of D has a multiplicative inverse.

If x_i is nonzero, we show that the set $x_i D = \{x_i x_0, x_i x_1, x_i x_2, \ldots, x_i x_n\}$ is the same as the set D. If $x_i x_j = x_i x_k$, then, by the cancellation property, $x_j = x_k$. Hence all the elements $x_i x_0, x_i x_1, x_i x_2, \ldots, x_i x_n$ are distinct, and $x_i D$ is a subset of D with the same number of elements. Therefore, $x_i D = D$. But then there is some element, x_j, such that $x_i x_j = x_1 = 1$. Hence $x_j = x_i^{-1}$, and D is a field. □

Note that $\mathbb{Z}$ is an *infinite* integral domain that is *not* a field.

Theorem 8.11. $\mathbb{Z}_n$ is a field if and only if n is prime.

Proof. Suppose that n is prime and that $[a] \cdot [b] = [0]$ in $\mathbb{Z}_n$. Then $n|ab$. So $n|a$ or $n|b$ by Euclid's Lemma (Theorem 12, Appendix 2). Hence $[a] = [0]$ or $[b] = [0]$, and $\mathbb{Z}_n$ is an integral domain. Since $\mathbb{Z}_n$ is also finite, it follows from Theorem 8.10 that $\mathbb{Z}_n$ is a field.

Suppose that n is not prime. Then we can write $n = rs$, where r and s are integers such that $1 < r < n$ and $1 < s < n$. Now $[r] \neq [0]$ and $[s] \neq [0]$ but $[r] \cdot [s] = [rs] = [0]$. Therefore, $\mathbb{Z}_n$ has zero divisors and hence is not a field. □

Example 8.12. Is $(\mathbb{Q}(\sqrt{2}), +, \cdot)$ an integral domain or a field?

Solution. From Example 8.3 we know that $\mathbb{Q}(\sqrt{2})$ is a commutative ring. Let $a + b\sqrt{2}$ be a nonzero element, so that at least one of a and b is not zero. Hence $a - b\sqrt{2} \neq 0$ (because $\sqrt{2}$ is not in $\mathbb{Q}$), so we have

$$\frac{1}{a + b\sqrt{2}} = \frac{a - b\sqrt{2}}{(a + b\sqrt{2})(a - b\sqrt{2})} = \frac{a}{a^2 - 2b^2} - \left(\frac{b}{a^2 - 2b^2}\right)\sqrt{2}.$$

This is an element of $\mathbb{Q}(\sqrt{2})$, and so is the inverse of $a + b\sqrt{2}$. Hence $\mathbb{Q}(\sqrt{2})$ is a field (and an integral domain). □

SUBRINGS AND MORPHISMS OF RINGS

If $(R, +, \cdot)$ is a ring, a nonempty subset S of R is called a **subring** of R if for all $a, b \in S$:

(i) $a + b \in S$.
(ii) $-a \in S$.
(iii) $a \cdot b \in S$.
(iv) $1 \in S$.

Conditions (i) and (ii) imply that $(S, +)$ is a subgroup of $(R, +)$ and can be replaced by the condition $a - b \in S$.

Proposition 8.13. If S is a subring of $(R, +, \cdot)$, then $(S, +, \cdot)$ is a ring.

Proof. Conditions (i) and (iii) of the definition above guarantee that S is closed under addition and multiplication. Condition (iv) shows that $1 \in S$. It follows from Proposition 3.8 that $(S, +)$ is a group. $(S, +, \cdot)$ satisfies the remaining axioms for a ring because they hold in $(R, +, \cdot)$. □

For example, $\mathbb{Z}, \mathbb{Q}$, and $\mathbb{R}$ are all subrings of $\mathbb{C}$. Let D be the set of $n \times n$ real diagonal matrices. Then D is a subring of the ring of all $n \times n$ real matrices, $M_n(\mathbb{R})$, because the sum, difference, and product of two diagonal matrices is another diagonal matrix. Note that D is commutative even though $M_n(\mathbb{R})$ is not.

Example 8.14. Show that $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} | a, b \in \mathbb{Q}\}$ is a subring of $\mathbb{R}$.

Solution. Let $a + b\sqrt{2}, c + d\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. Then

(i) $(a + b\sqrt{2}) + (c + d\sqrt{2}) = (a + c) + (b + d)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.
(ii) $-(a + b\sqrt{2}) = (-a) + (-b)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.
(iii) $(a + b\sqrt{2}) \cdot (c + d\sqrt{2}) = (ac + 2bd) + (ad + bc)\sqrt{2} \in \mathbb{Q}(\sqrt{2})$.
(iv) $1 = 1 + 0\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. □

A morphism between two rings is a function between their underlying sets that preserves the two operations of addition and multiplication and also the element 1. Many authors use the term *homomorphism* instead of *morphism*.

More precisely, let $(R, +, \cdot)$ and $(S, +, \cdot)$ be two rings. The function $f: R \to S$ is called a **ring morphism** if for all $a, b \in R$:

(i) $f(a + b) = f(a) + f(b)$.
(ii) $f(a \cdot b) = f(a) \cdot f(b)$.
(iii) $f(1) = 1$.

If the operations in the two rings are denoted by different symbols, for example, if the rings are $(R, +, \cdot)$ and $(S, \oplus, \circ)$, then the conditions for $f: R \to S$ to be a ring morphism are:

(i) $f(a + b) = f(a) \oplus f(b)$.
(ii) $f(a \cdot b) = f(a) \circ f(b)$.
(iii) $f(1_R) = 1_S$ where 1_R and 1_S are the respective identities.

A **ring isomorphism** is a bijective ring morphism. If there is an isomorphism between the rings R and S, we say R and S are **isomorphic rings** and write $R \cong S$.

A ring morphism, f, from $(R, +, \cdot)$ to $(S, +, \cdot)$ is, in particular, a group morphism from $(R, +)$ to $(S, +)$. Therefore, by Proposition 3.19, $f(0) = 0$ and $f(-a) = -f(a)$ for all $a \in R$.

The inclusion function, $i: S \to R$, of any subring S into a ring R is always a ring morphism. The function $f: \mathbb{Z} \to \mathbb{Z}_n$, defined by $f(x) = [x]$, which maps an integer to its equivalence class modulo n, is a ring morphism from $(\mathbb{Z}, +, \cdot)$ to $(\mathbb{Z}_n, +, \cdot)$.

Example 8.15. If X is a one element set, show that $f: \mathcal{P}(X) \to \mathbb{Z}_2$ is a ring isomorphism between $(\mathcal{P}(X), \Delta, \cap)$ and $(\mathbb{Z}_2, +, \cdot)$, where $f(\emptyset) = [0]$ and $f(X) = [1]$.

Solution. We can check that f is a morphism by testing all the possibilities for $f(A \Delta B)$ and $f(A \cap B)$. Since the rings are commutative, they are

$$\begin{aligned}
f(\emptyset \Delta \emptyset) &= f(\emptyset) = [0] = f(\emptyset) + f(\emptyset) \\
f(\emptyset \Delta X) &= f(X) = [1] = f(\emptyset) + f(X) \\
f(X \Delta X) &= f(\emptyset) = [0] = f(X) + f(X) \\
f(\emptyset \cap \emptyset) &= f(\emptyset) = [0] = f(\emptyset) \cdot f(\emptyset) \\
f(\emptyset \cap X) &= f(\emptyset) = [0] = f(\emptyset) \cdot f(X) \\
f(X \cap X) &= f(X) = [1] = f(X) \cdot f(X).
\end{aligned}$$

Both rings contain only two elements, and f is a bijection; therefore, f is an isomorphism. □

If $f: R \to S$ is an isomorphism between two finite rings, the addition and multiplication tables of S will be the same as those of R if we replace each $a \in R$ by $f(a) \in S$. For example, Tables 8.4 and 8.5 illustrate the isomorphism of Example 8.15.

The following ring isomorphism between linear transformations and matrices is the crux of much of linear algebra.

Example 8.16. The linear transformations from $\mathbb{R}^n$ to itself form a ring, $(\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n), +, \circ)$, under addition and composition. Show that the function

$$f: \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n) \to M_n(\mathbb{R})$$

TABLE 8.4. Ring $\mathcal{P}(X)$ When X Is a Point

Δ	$\emptyset$	X
$\emptyset$	$\emptyset$	X
X	X	$\emptyset$

$\cap$	$\emptyset$	X
$\emptyset$	$\emptyset$	$\emptyset$
X	$\emptyset$	X

TABLE 8.5. Ring $\mathbb{Z}_2$

$+$	[0]	[1]
[0]	[0]	[1]
[1]	[1]	[0]

$\cdot$	[0]	[1]
[0]	[0]	[0]
[1]	[0]	[1]

is a ring morphism, where f assigns to each linear transformation its standard matrix, that is, its $n \times n$ coefficient matrix with respect to the standard basis of $\mathbb{R}^n$.

Solution. If α is a linear transformation from $\mathbb{R}^n$ to itself, then

$$\alpha \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \qquad\qquad \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n \end{bmatrix} \text{ and } f(\alpha) = \begin{bmatrix} a_{11} \ldots a_{1n} \\ \vdots \qquad \vdots \\ a_{n1} \ldots a_{nn} \end{bmatrix}.$$

Matrix addition and multiplication is defined so that $f(\alpha + \beta) = f(\alpha) + f(\beta)$ and $f(\alpha \circ \beta) = f(\alpha) \cdot f(\beta)$. Also, if ι is the identity linear transformation then $f(\iota)$ is the identity matrix.

Any matrix defines a linear transformation, so that f is surjective. Furthermore, f is injective, because any matrix can arise from only one linear transformation. In fact, the jth column of the matrix must be the image of the jth basis vector. Hence f is an isomorphism. □

Example 8.17. Show that $f: \mathbb{Z}_{24} \to \mathbb{Z}_4$, defined by $f([x]_{24}) = [x]_4$ is a ring morphism.

Proof. Since the function is defined in terms of representatives of equivalence classes, we first check that it is well defined. If $[x]_{24} = [y]_{24}$, then $x \equiv y \bmod 24$ and $24|(x - y)$. Hence $4|(x - y)$ and $[x]_4 = [y]_4$, which shows that f is well defined.

We now check the conditions for f to be a ring morphism.

(i) $f([x]_{24} + [y]_{24}) = f([x + y]_{24}) = [x + y]_4 = [x]_4 + [y]_4$.
(ii) $f([x]_{24} \cdot [y]_{24}) = f([xy]_{24}) = [xy]_4 = [x]_4 \cdot [y]_4$.
(iii) $f([1]_{24}) = [1]_4$. □

NEW RINGS FROM OLD

This section introduces various methods for constructing new rings from given rings. These include the direct product of rings, matrix rings, polynomial rings, rings of sequences, and rings of formal power series. Perhaps the most important class of rings constructible from given rings is the class of quotient rings. Their construction is analogous to that of quotient groups and is discussed in Chapter 10.

If $(R, +, \cdot)$ and $(S, +, \cdot)$ are two rings, their **product** is the ring $(R \times S, +, \cdot)$ whose underlying set is the cartesian product of R and S and whose operations are defined component-wise by

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2) \quad \text{and} \quad (r_1, s_1) \cdot (r_2, s_2) = (r_1 \cdot r_2, s_1 \cdot s_2).$$

It is readily verified that these operations do indeed define a ring structure on $R \times S$ whose zero is $(0_R, 0_S)$, where 0_R and 0_S are the zeros of R and S, and whose multiplicative identity is $(1_R, 1_S)$, where 1_R and 1_S are the identities in R and S.

The product construction can be iterated any number of times. For example, $(\mathbb{R}^n, +, \cdot)$ is a commutative ring, where $\mathbb{R}^n$ is the n-fold cartesian product of $\mathbb{R}$ with itself.

Example 8.18. Write down the addition and multiplication tables for $\mathbb{Z}_2 \times \mathbb{Z}_3$.

Solution. Let $\mathbb{Z}_2 = \{0, 1\}$ and $\mathbb{Z}_3 = \{0, 1, 2\}$. Then $\mathbb{Z}_2 \times \mathbb{Z}_3 = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2)\}$. The addition and multiplication tables are given in Table 8.6. In calculating these, it must be remembered that addition and multiplication are performed modulo 2 in the first coordinate and modulo 3 in the second coordinate. □

We know that $\mathbb{Z}_2 \times \mathbb{Z}_3$ and $\mathbb{Z}_6$ are isomorphic as *groups*; we now show that they are isomorphic as *rings*.

Theorem 8.19. $\mathbb{Z}_m \times \mathbb{Z}_n$ is isomorphic as a ring to $\mathbb{Z}_{mn}$ if and only if $\gcd(m, n) = 1$.

Proof. If $\gcd(m, n) = 1$, it follows from Theorems 4.32 and 3.20 that the function

$$f: \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n$$

TABLE 8.6. Ring $\mathbb{Z}_2 \times \mathbb{Z}_3$

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

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

defined by $f([x]_{mn}) = ([x]_m, [x]_n)$ is a group isomorphism. However, this function also preserves multiplication because

$$f([x]_{mn} \cdot [y]_{mn}) = f([xy]_{mn}) = ([xy]_m, [xy]_n) = ([x]_m[y]_m, [x]_n[y]_n)$$
$$= ([x]_m, [x]_n) \cdot ([y]_m, [y]_n) = f([x]_{mn}) \cdot f([y]_{mn}).$$

Also, $f([1]_{mn}) = ([1]_m, [1]_n)$; thus f is a ring isomorphism.

It was shown in the discussion following Corollary 4.33 that if $\gcd(m, n) \neq 1$, $\mathbb{Z}_m \times \mathbb{Z}_n$ and $\mathbb{Z}_{mn}$ are not isomorphic as groups, and hence they cannot be isomorphic as rings. □

We can extend this result by induction to show the following.

Theorem 8.20. Let $m = m_1 \cdot m_2 \cdots m_r$, where $\gcd(m_i, m_j) = 1$ if $i \neq j$. Then $\mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \cdots \times \mathbb{Z}_{m_r}$ is a ring isomorphic to $\mathbb{Z}_m$.

Corollary 8.21. Let $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$ be a decomposition of the integer n into powers of distinct primes. Then $\mathbb{Z}_n \cong \mathbb{Z}_{p_1^{\alpha_1}} \times \mathbb{Z}_{p_2^{\alpha_2}} \times \cdots \times \mathbb{Z}_{p_r^{\alpha_r}}$ as rings.

If R is a commutative ring, we can construct the **ring of $n \times n$ matrices with entries from R**, $(M_n(R), +, \cdot)$. Addition and multiplication are performed as in real matrices.

For example, $(M_n(\mathbb{Z}_2), +, \cdot)$ is the ring of $n \times n$ matrices with 0 and 1 entries. Addition and multiplication is performed modulo 2. This is a noncommutative ring with $2^{(n^2)}$ elements.

If R is a commutative ring, a **polynomial** $p(x)$ in the indeterminate x over the ring R is an expression of the form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

where $a_0, a_1, a_2, \ldots, a_n \in R$ and $n \in \mathbb{N}$. The element a_i is called the **coefficient** of x^i in $p(x)$. If the coefficient of x^i is zero, the term $0x^i$ may be omitted, and if the coefficient of x^i is one, $1x^i$ may be written simply as x^i.

Two polynomials $f(x)$ and $g(x)$ are called **equal** when they are identical, that is, when the coefficient of x^n is the same in each polynomial for every $n \geqslant 0$. In particular,

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = 0$$

is the zero polynomial if and only if $a_0 = a_1 = a_2 = \cdots = a_n = 0$.

If n is the largest integer for which $a_n \neq 0$, we say that $p(x)$ has **degree** n and write $\deg p(x) = n$. If all the coefficients of $p(x)$ are zero, then $p(x)$ is called the **zero polynomial**, and its degree is not defined.

For example, $4x^2 - \sqrt{3}$ is a polynomial over $\mathbb{R}$ of degree 2, $ix^4 - (2 + i)x^3 + 3x$ is a polynomial over $\mathbb{C}$ of degree 4, and $x^7 + x^5 + x^4 + 1$ is a polynomial over $\mathbb{Z}_2$ of degree 7. The number 5 is a polynomial over $\mathbb{Z}$ of degree 0; the zero

polynomial and the polynomials of degree 0 are called **constant polynomials** because they contain no x terms.

The set of all polynomials in x with coefficients from the commutative ring R is denoted by $R[x]$. That is,

$$R[x] = \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n | a_i \in R, n \in \mathbb{N}\}.$$

This forms a ring $(R[x], +, \cdot)$ called the **polynomial ring with coefficients from** R when addition and multiplication of the polynomials

$$p(x) = \sum_{i=0}^{n} a_i x^i \quad \text{and} \quad q(x) = \sum_{i=0}^{m} b_i x^i$$

are defined by

$$p(x) + q(x) = \sum_{i=0}^{\max(m,n)} (a_i + b_i)x^i$$

and

$$p(x) \cdot q(x) = \sum_{k=0}^{m+n} c_k x^k \quad \text{where} \quad c_k = \sum_{i+j=k} a_i b_j.$$

With a little effort, it can be verified that $(R[x], +, \cdot)$ satisfies all the axioms for a commutative ring. The zero is the zero polynomial, and the multiplicative identity is the constant polynomial 1.

For example, in $\mathbb{Z}_5[x]$, the polynomial ring with coefficients in the integers modulo 5, we have

$$(2x^3 + 2x^2 + 1) + (3x^2 + 4x + 1) = 2x^3 + 4x + 2$$

and

$$(2x^3 + 2x^2 + 1) \cdot (3x^2 + 4x + 1) = x^5 + 4x^4 + 4x + 1.$$

When working in $\mathbb{Z}_n[x]$, the coefficients, but *not* the exponents, are reduced modulo n.

Proposition 8.22. If R is an integral domain and $p(x)$ and $q(x)$ are nonzero polynomials in $R[x]$, then

$$\deg(p(x) \cdot q(x)) = \deg\ p(x) + \deg\ q(x).$$

Proof. Let $\deg\ p(x) = n$, $\deg\ q(x) = m$ and let $p(x) = a_0 + \cdots + a_nx^n$, $q(x) = b_0 + \cdots + b_mx^m$, where $a_n \neq 0, b_m \neq 0$. Then the coefficient of the highest power of x in $p(x) \cdot q(x)$ is a_nb_m, which is nonzero since R has no zero divisors. Hence $\deg(p(x) \cdot q(x)) = m + n$. □

If the coefficient ring is not an integral domain, the degree of a product may be less than the sum of the degrees. For example, $(2x^3 + x) \cdot (3x) = 3x^2$ in $\mathbb{Z}_6[x]$.

Corollary 8.23. If R is an integral domain, so is $R[x]$.

Proof. If $p(x)$ and $q(x)$ are nonzero elements of $R[x]$, then $p(x) \cdot q(x)$ is also nonzero by Proposition 8.22. Hence $R[x]$ has no zero divisors. □

The construction of a polynomial ring can be iterated to obtain the ring of polynomials in n variables $x_1, \ldots, x_n$, with coefficients from R. We define inductively $R[x_1, \ldots, x_n] = R[x_1, \ldots, x_{n-1}][x_n]$. For example, consider a polynomial f in $R[x, y] = R[x][y]$, say

$$f = f_0 + f_1 y + f_2 y^2 + \cdots + f_n y^n,$$

where each $f_i = f_i(x)$ is in $R[x]$. If we write $f_i = a_{0i} + a_{1i}x + a_{2i}x^2 \cdots$ for each i, then

$$f(x, y) = a_{00} + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 + \cdots.$$

Clearly, we can prove by induction from Corollary 8.22 that $R[x_1, \ldots, x_n]$ is an integral domain if R is an integral domain.

Proposition 8.24. Let R be a commutative ring and denote the infinite sequence of elements of R, $\langle a_0, a_1, a_2, \ldots \rangle$, by $\langle a_i \rangle$. Define addition, $+$, and convolution, $*$, of two such sequences by

$$\langle a_i \rangle + \langle b_i \rangle = \langle a_i + b_i \rangle$$

and

$$\langle a_i \rangle * \langle b_i \rangle = \left\langle \sum_{j+k=i} a_j b_k \right\rangle = \langle a_0 b_i + a_1 b_{i-1} + \cdots + a_i b_0 \rangle.$$

The set of all such sequences forms a commutative ring $(R^{\mathbb{N}}, +, *)$ called the **ring of sequences** in R. If R is an integral domain, so is $R^{\mathbb{N}}$.

Proof. Addition is clearly associative and commutative. The zero element is the zero sequence $\langle 0 \rangle = \langle 0, 0, 0, \ldots \rangle$, and the negative of $\langle a_i \rangle$ is $\langle -a_i \rangle$. Now

$$\begin{aligned}(\langle a_i \rangle * \langle b_i \rangle) * \langle c_i \rangle &= \left\langle \sum_{j+k=i} a_j b_k \right\rangle * \langle c_i \rangle \\ &= \left\langle \sum_{l+m=i} \left(\sum_{j+k=m} a_j b_k \right) c_l \right\rangle = \left\langle \sum_{j+k+l=i} a_j b_k c_l \right\rangle.\end{aligned}$$

Similarly, bracketing the sequences in the other way, we obtain the same result, which shows that convolution is associative.

Convolution is clearly commutative and the distributive laws hold because

$$\langle a_i\rangle * (\langle b_i\rangle + \langle c_i\rangle) = \left\langle \sum_{j+k=i} a_j(b_k + c_k) \right\rangle$$
$$= \left\langle \sum_{j+k=i} a_j b_k \right\rangle + \left\langle \sum_{j+k=i} a_j c_k \right\rangle = \langle a_i\rangle * \langle b_i\rangle + \langle a_i\rangle * \langle c_i\rangle.$$

The identity in the ring of sequences is $\langle 1, 0, 0, \ldots\rangle$ because

$$\langle 1, 0, 0, \ldots\rangle * \langle a_0, a_1, a_2, \ldots\rangle = \langle 1a_0, 1a_1 + 0a_0, 1a_2 + 0a_1 + 0a_0, \ldots\rangle$$
$$= \langle a_0, a_1, a_2, \ldots\rangle.$$

Therefore, $(R^{\mathbb{N}}, +, *)$ is a commutative ring.

Suppose that a_q and b_r are the first nonzero elements in the nonzero sequences $\langle a_i\rangle$ and $\langle b_i\rangle$, respectively. Then the element in the $(q + r)$th position of their convolution is

$$\sum_{j+k=q+r} a_j b_k = a_0 b_{q+r} + a_1 b_{q+r-1} + \cdots + a_q b_r + a_{q+1} b_{r-1} + \cdots + a_{q+r} b$$
$$= 0 \quad + \quad 0 \quad + \cdots + a_q b_r + \quad 0 \quad + \cdots + 0 = a_q b_r.$$

Hence, if R is an integral domain, this element is not zero and the ring of sequences has no zero divisors. □

The ring of sequences cannot be a field because $\langle 0, 1, 0, 0, \ldots\rangle$ has no inverse. In fact, for any sequence $\langle b_i\rangle$, $\langle 0, 1, 0, 0, \ldots\rangle * \langle b_0, b_1, b_2, \ldots\rangle = \langle 0, b_0, b_1, \ldots\rangle$, which can never be the identity in the ring.

A **formal power series** in x with coefficients from a commutative ring R is an expression of the form.

$$a_0 + a_1 x + a_2 x^2 + \cdots = \sum_{i=0}^{\infty} a_i x^i \quad \text{where} \quad a_i \in R.$$

In contrast to a polynomial, these power series can have an infinite number of nonzero terms.

We denote the set of all such formal power series by $R[[x]]$. The term *formal* is used to indicate that questions of convergence of these series are not considered. Indeed, over many rings, such as $\mathbb{Z}_n$, convergence would not be meaningful.

Motivated by $\mathbb{R}^{\mathbb{N}}$, addition and multiplication are defined in $R[[x]]$ by

$$\left(\sum_{i=0}^{\infty} a_i x^i\right) + \left(\sum_{i=0}^{\infty} b_i x^i\right) = \sum_{i=0}^{\infty} (a_i + b_i) x^i$$

and

$$\left(\sum_{i=0}^{\infty} a_i x^i\right) \cdot \left(\sum_{i=0}^{\infty} b_i x^i\right) = \sum_{i=0}^{\infty} \left(\sum_{j+k=i} a_j b_k\right) x^i.$$

It can be verified that these formal power series do form a ring, $(R[[x]], +, \cdot)$, and that the polynomial ring, $R[x]$, is the subring consisting of those power series with only a finite number of nonzero terms. In fact, the ring of sequences $(R^{\mathbb{N}}, +, *)$ is isomorphic to the ring of formal power series $(R[[x]], +, \cdot)$. The function $f: R^{\mathbb{N}} \to R[[x]]$ that is defined by $f(\langle a_0, a_1, a_2, \cdots\rangle) = a_0 + a_1 x + a_2 x^2 + \cdots$ is clearly a bijection. It follows from the definitions of addition, multiplication, and convolution in these rings, that f is a ring morphism.

FIELD OF FRACTIONS

We can always add, subtract, and multiply elements in any ring, but we cannot always divide. However, if the ring is an integral domain, it is possible to enlarge it so that division by nonzero elements is possible. In other words, we can construct a field containing the given ring as a subring. This is precisely what we did following Example 4.2 when constructing the rational numbers from the integers.

If the original ring did have zero divisors or was noncommutative, it could not possibly be a subring of any field, because fields cannot contain zero divisors or pairs of noncommutative elements.

Theorem 8.25. If R is an integral domain, it is possible to construct a field Q, so that the following hold:

(i) R is isomorphic to a subring, R', of Q.
(ii) Every element of Q can be written as $p \cdot q^{-1}$ for suitable $p, q \in R'$.

Q is called the **field of fractions** of R (or sometimes the **field of quotients** of R).

Proof. Consider the set $R \times R^* = \{(a, b) | a, b \in R, b \neq 0\}$, consisting of pairs of elements of R, the second being nonzero. Motivated by the fact that $\frac{a}{b} = \frac{c}{d}$ in $\mathbb{Q}$ if and only if $ad = bc$, we define a relation $\sim$ on $R \times R^*$ by

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

We verify that this is an equivalence relation.

(i) $(a, b) \sim (a, b)$, since $ab = ba$.
(ii) If $(a, b) \sim (c, d)$, then $ad = bc$. This implies that $cb = da$ and hence that $(c, d) \sim (a, b)$.

(iii) If $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$, then $ad = bc$ and $cf = de$. This implies that $(af - be)d = (ad)f - b(ed) = bcf - bcf = 0$. Since R has no zero divisors and $d \neq 0$, it follows that $af = be$ and $(a, b) \sim (e, f)$.

Hence the relation $\sim$ is reflexive, symmetric, and transitive.

Denote the equivalence class containing (a, b) by a/b and the set of equivalence classes by Q. As in $\mathbb{Q}$, define addition and multiplication in Q by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

These operations on equivalence classes are defined in terms of particular representatives, so it must be checked that they are well defined. If $a/b = a'/b'$ and $c/d = c'/d'$, then $ab' = a'b$ and $cd' = c'd$. Hence

$$\begin{aligned}(ad + bc)(b'd') &= (ab')dd' + bb'(cd') = (a'b)dd' + bb'(c'd) \\ &= (a'd' + b'c')(bd)\end{aligned}$$

and therefore $\dfrac{ad + bc}{bd} = \dfrac{a'd' + b'c'}{b'd'}$, which shows that addition is well defined. Also, $acb'd' = a'c'bd$; thus $\dfrac{ac}{bd} = \dfrac{a'c'}{b'd'}$, which shows that multiplication is well defined.

It can now be verified that $(Q, +, \cdot)$ is a field. The zero is 0/1, and the identity is 1/1. For example, the distributive laws hold because

$$\begin{aligned}\frac{a}{b} \cdot \left(\frac{c}{d} + \frac{e}{f}\right) &= \frac{a}{b} \cdot \frac{cf + de}{df} = \frac{a(cf + de)}{bdf} = \frac{a(cf + de)}{bdf} \cdot \frac{b}{b} = \frac{ac}{bd} + \frac{ae}{bf} \\ &= \frac{a}{b} \cdot \frac{c}{d} + \frac{a}{b} \cdot \frac{e}{f}.\end{aligned}$$

The inverse of any nonzero element $\dfrac{a}{b}$ is $\dfrac{b}{a}$. The remaining axioms for a field are straightforward to check.

The ring R is isomorphic to the subring $R' = \left\{ \dfrac{r}{1} \middle| r \in R \right\}$ of Q by an isomorphism that maps r to $\dfrac{r}{1}$. Any element $\dfrac{a}{b}$ in the field Q can be written as $\dfrac{a}{b} = \dfrac{a}{1} \cdot \dfrac{1}{b} = \dfrac{a}{1}\left(\dfrac{b}{1}\right)^{-1}$ where $\dfrac{a}{1}$ and $\dfrac{b}{1}$ are in R'. □

If we take $R = \mathbb{Z}$ to be integers in the above construction, we obtain the rational numbers $\mathbb{Q}$ as the field of fractions.

If R is an integral domain, the field of fractions of the polynomial ring $R[x]$ is called the *field of* **rational functions with coefficients in** R. Its elements can be considered as fractions of one polynomial over a nonzero polynomial.

A (possibly noncommutative) ring is called a **domain** if $ab = 0$ if and only if $a = 0$ or $b = 0$. Thus the commutative domains are precisely the integral domains. In 1931, Oystein Ore (1899–1968) extended Theorem 8.25 to a class of domains (now called **left Ore domains**) for which a ring of *left fractions* can be constructed that is a **skew field** (that is, a field that is not necessarily commutative). On the other hand, in 1937, A. I. Mal'cev (1909–1967) discovered an example of a domain that cannot be embedded in *any* skew field. The simplest example of a noncommutative skew field is the ring of **quaternions** (see Exercise 8.36). It has infinitely many elements, in agreement with a famous theorem of J. H. M. Wedderburn, proved in 1905, asserting that any finite skew field is necessarily commutative.

CONVOLUTION FRACTIONS

We now present an application of the field of fractions that has important implications in analysis. This example is of a different type than most of the applications in this book. It can be omitted, without loss of continuity, by those readers not interested in analysis or applied mathematics.

We construct the field of fractions of a set of continuous functions, and use it to explain two mathematical techniques that have been used successfully by engineers and physicists for many years, but were at first mistrusted by mathematicians because they did not have a firm mathematical basis. One such technique was introduced by O. Heaviside in 1893 in dealing with electrical circuits; this is called the **operational calculus**, and it enabled him to solve partial differential equations by manipulating differential operators as if they were algebraic quantities. The second such technique is the use of **impulse functions** in applied mathematics and mathematical physics. In 1926, when solving problems in relativistic quantum mechanics, P. Dirac introduced his **delta function**, $\delta(x)$, which has the property that

$$\delta(x) = 0 \quad \text{if } x \neq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(x)\, dx = 1.$$

If we use the usual definition of functions, no such object exists. However, it can be pictured in Figure 8.1 as the limit, as k tends to zero, of the functions $\delta_k(x)$, where

$$\delta_k(x) = \begin{cases} 1/k & \text{if } 0 \leqslant x \leqslant k. \\ 0 & \text{otherwise} \end{cases}$$

Each function $\delta_k(x)$ vanishes outside the interval $0 \leqslant x \leqslant k$ and has the property that

$$\int_{-\infty}^{\infty} \delta_k(x)\, dx = 1.$$

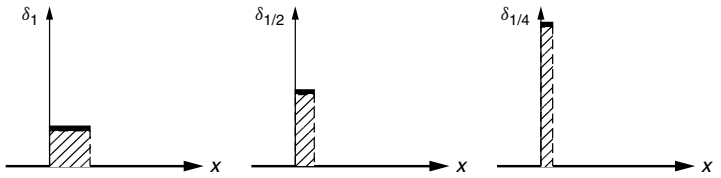


Figure 8.1. The Dirac delta "function" is the limit of δ_k as k tends to zero.

Consider the set, $C[0, \infty)$, of real-valued functions that are continuous in the interval $0 \leqslant x < \infty$. We define the operations of addition and convolution on this set so that the algebraic structure $(C[0, \infty), +, *)$ is *nearly* an integral domain: convolution does not have an identity so the structure fails to satisfy Axiom (vi) of a ring. However, it is still possible to embed this structure in its field of fractions. The Polish mathematician Jan Mikusinski constructed this field of fractions and called such elements *operators* or *generalized functions*. The Dirac delta function is a generalized function and is in fact the identity for convolution in the field of fractions.

Define **addition** and **convolution** of two functions f and g in $C[0, \infty)$ by

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (f * g)(x) = \int_0^x f(t)g(x - t)\,dt.$$

This convolution of functions is the continuous analogue of convolution of sequences, as can be seen by writing the ith term of the sequence

$$\langle a_i \rangle * \langle b_i \rangle \quad \text{as} \quad \sum_{t=0}^{i} a_t b_{i-t}.$$

It is clear that addition is associative and commutative, and the zero function is the additive identity. Also, the negative of $f(x)$ is $-f(x)$.

Convolution is commutative because

$$\begin{aligned}(f * g)(x) &= \int_0^x f(t)g(x - t)\,dt \\ &= -\int_x^0 f(x - u)g(u)\,du \qquad \text{substituting } u = x - t \\ &= \int_0^x g(u)f(x - u)\,du = (g * f)(x).\end{aligned}$$

Convolution is associative because

$$\begin{aligned}(f * (g * h))(x) &= \int_0^x f(t)(g * h)(x - t)\,dt \\ &= \int_0^x f(t)\left[\int_0^{x-t} g(u)h(x - t - u)\,du\right] dt \\ &= \int_0^x f(t)\left[\int_t^x g(w - t)h(x - w)\,dw\right] dt,\end{aligned}$$

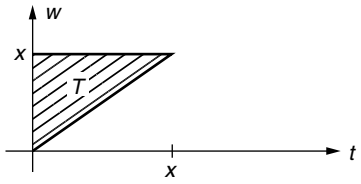


Figure 8.2

putting $u = w - t$. This integration is over the triangle in Figure 8.2 so, changing the order of integration,

$$\begin{aligned}(f * (g * h))(x) &= \int_0^x \int_0^w f(t)g(w-t)h(x-w)\,dt\,dw, \\ &= \int_0^x (f * g)(w)h(x-w)\,dw = ((f * g) * h)(x).\end{aligned}$$

The distributive laws follow because

$$\begin{aligned}((f+g) * h)(x) &= \int_0^x (f(t) + g(t))h(x-t)\,dt \\ &= \int_0^x f(t)h(x-t)\,dt + \int_0^x g(t)h(x-t)\,dt \\ &= (f * h)(x) + (g * h)(x).\end{aligned}$$

If f is a function that is the identity under convolution, then $f * h = h$ for all functions h. If we take h to be the function defined by $h(x) = 1$ for all $0 \leqslant x < \infty$, then

$$(f * h)(x) = \int_0^x f(t)\,dt = 1 \quad \text{for all } x \geqslant 0.$$

There is no function f in $C[0, \infty)$ with this property, although the Dirac delta "function" does have this property. Hence $(C[0, \infty), +, *)$ satisfies all the axioms for a commutative ring except for the existence of an identity under convolution.

Furthermore, there are no zero divisors under convolution; that is, $f * g = 0$ implies that $f = 0$ or $g = 0$. This is a hard result in analysis, which is known as *Titchmarsh's theorem*. Proofs can be found in Erdelyi [36] or Marchand [37].

However, we can still construct the field of fractions of this algebraic object in exactly the same way as we did in Theorem 8.25. For example, the even integers under addition and multiplication, $(2\mathbb{Z}, +, \cdot)$ is also an algebraic object that satisfies all the axioms for an integral domain except for the fact that multiplication has no identity. The field of fractions of $2\mathbb{Z}$ is the set of rational numbers; every rational number can be written in the form $2r/2s$, where $2r, 2s \in 2\mathbb{Z}$.

The field of fractions of $(C[0, \infty), +, *)$ is called the **field of convolution fractions**, and its elements are sometimes called **generalized functions, distributions**, or **operators**. Elements of this field are the abstract entities f/g, where f and g are functions. There is a bijection between the set of elements of the form $f * g/g$ and the set $C[0, \infty)$. It is possible to interpret other convolution fractions as impulse functions, discontinuous functions, and even differential or integral operators. The Dirac delta function can be defined to be the identity of this field under convolution; therefore, $\delta = f/f$, for any nonzero function f.

The Heaviside step function illustrated in Figure 8.3 is defined by $h(x) = 1$ if $x \geqslant 0$, and $h(x) = 0$ if $x < 0$. The function is continuous when restricted to the nonnegative numbers and, in some sense, is the integral of the Dirac delta function. Convolution by h acts as an integral operator on any continuous function because

$$(h * f)(x) = (f * h)(x) = \int_0^x f(t)h(x - t)\,dt = \int_0^x f(t)\,dt.$$

Hence $h * f$ is the integral of f. We can use this to define integration of any generalized function. Take the integral of the convolution fraction f/g to be the fraction $(h * f)/g$.

Denote the inverse of the Heaviside step function by s, so that $s = h/h * h$. This element s is not a genuine function, but only a convolution fraction. Convolution by s acts, in some sense, as a differential operator in the field of convolution fractions. It is not exactly the usual differential operator, because convolution by s and by h must commute, and $s * h = h * s$ must be the identity. If $f(x)$ is a continuous function, we know from the calculus, that the derivative of $\int_0^x f(t)\,dt$ is just $f(x)$; however, $\int_0^x f'(t)\,dt$ is not just $f(x)$ but is $f(x) - f(0)$. In fact, if the function f has a derivative,

$$(s * f)(x) = f'(x) + f(0)\delta(x)$$

where $\delta(x)$ is the identity in the field of convolution fractions. Now, when we calculate $h * s * f$, which is equivalent to integrating $s * f$ from 0 to x, we obtain the function f back again.

By repeated convolution with s or h, a generalized function can be differentiated or integrated any number of times, the result being another generalized function. We can even differentiate or integrate a fractional number of times. These

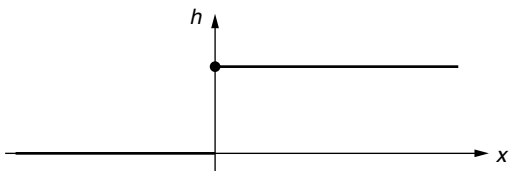


Figure 8.3. Heaviside step function.

operations s and h can be used to explain Heaviside's operational calculus, in which differential and integral operators are manipulated like algebraic symbols.

For further information on the algebraic aspects of generalized functions and distributions, see Erdelyi [36] or Marchand [37].

EXERCISES

8.1. Write out the tables for the ring $\mathbb{Z}_4$.

8.2. Write out the tables for the ring $\mathbb{Z}_2 \times \mathbb{Z}_2$.

Which of the systems described in Exercises 8.3 to 8.12 are rings under addition and multiplication? Give reasons.

8.3. $\{a + b\sqrt{5} | a, b \in \mathbb{Z}\}$.

8.4. $\mathbb{N}$.

8.5. $\{a + b\sqrt{2} + c\sqrt{3} | a, b, c \in \mathbb{Z}\}$.

8.6. $\{a + \sqrt[3]{2}b | a, b \in \mathbb{Q}\}$.

8.7. All 2×2 real matrices with zero determinant.

8.8. All rational numbers that can be written with denominator 2.

8.9. All rational numbers that can be written with an odd denominator.

8.10. $(\mathbb{Z}, +, \times)$, where $+$ is the usual addition and $a \times b = 0$ for all $a, b \in \mathbb{Z}$.

8.11. The set $A = \{a, b, c\}$ with tables given in Table 8.7.

TABLE 8.7

$+$	a	b	c
a	a	b	c
b	b	c	a
c	c	a	b

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

8.12. The set $A = \{a, b, c\}$ with tables given in Table 8.8.

TABLE 8.8

$+$	a	b	c
a	a	b	c
b	b	c	a
c	c	a	b

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

8.13. A ring R is called a **boolean ring** if $a^2 = a$ for all $a \in R$.

(**a**) Show that $(\mathcal{P}(X), \Delta, \cap)$ is a boolean ring for any set X.

(**b**) Show that $\mathbb{Z}_2$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are boolean rings.

(**c**) Prove that if R is boolean, then $2a = 0$ for all $a \in R$.

(**d**) Prove that any boolean ring is commutative.

(**e**) If $(R, \wedge, \vee, ')$ is any boolean algebra, show that $(R, \Delta, \wedge)$ is a boolean ring where $a \Delta b = (a \wedge b') \vee (a' \wedge b)$.

(**f**) If $(R, +, \cdot)$ is a boolean ring, show that $(R, \wedge, \vee, ')$ is a boolean algebra where $a \wedge b = a \cdot b$, $a \vee b = a + b + a \cdot b$ and $a' = 1 + a$.

This shows that there is a one-to-one correspondence between boolean algebras and boolean rings.

8.14. If A and B are subrings of a ring R, prove that $A \cap B$ is also a subring of R.

8.15. Prove that the only subring of $\mathbb{Z}_n$ is itself.

Which of the sets described in Exercises 8.16 to 8.20 are subrings of $\mathbb{C}$? Give reasons.

8.16. $\{0 + ib | b \in \mathbb{R}\}$.

8.17. $\{a + ib | a, b \in \mathbb{Q}\}$.

8.18. $\{a + b\sqrt{-7} | a, b \in \mathbb{Z}\}$.

8.19. $\{z \in \mathbb{C} | \, |z| \leqslant 1\}$.

8.20. $\{a + ib | a, b \in \mathbb{Z}\}$.

Which of the rings described in Exercises 8.21 to 8.26 are integral domains and which are fields?

8.21. $\mathbb{Z}_2 \times \mathbb{Z}_2$.

8.22. $(\mathcal{P}(\{a\}), \Delta, \cap)$.

8.23. $\{a + bi | a, b \in \mathbb{Q}\}$.

8.24. $\mathbb{Z} \times \mathbb{R}$.

8.25. $\{a + b\sqrt{2} | a, b \in \mathbb{Z}\}$.

8.26. $\mathbb{R}[x]$.

8.27. Prove that the set $C(\mathbb{R})$ of continuous real-valued functions defined on the real line forms a ring $(C(\mathbb{R}), +, \cdot)$, where addition and multiplication of two functions $f, g \in C(\mathbb{R})$ is given by

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (f \cdot g)(x) = f(x) \cdot g(x).$$

Find all the zero divisors in the rings described in Exercises 8.28 to 8.33.

8.28. $\mathbb{Z}_4$.

8.29. $\mathbb{Z}_{10}$.

8.30. $\mathbb{Z}_4 \times \mathbb{Z}_2$.

8.31. $(\mathcal{P}(X), \Delta, \cap)$.

8.32. $M_2(\mathbb{Z}_2)$.

8.33. $M_n(\mathbb{R})$.

8.34. Let $(R, +, \cdot)$ be a ring in which $(R, +)$ is a cyclic group. Prove that $(R, +, \cdot)$ is commutative ring.

8.35. Show that $S = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \middle| a, b \in \mathbb{R} \right\}$ is a subring of $M_2(\mathbb{R})$ isomorphic to $\mathbb{C}$.

8.36. Show that $\mathbb{H} = \left\{ \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \middle| \alpha, \beta \in \mathbb{C} \right\}$ is a subring of $M_2(\mathbb{C})$, where $\overline{\alpha}$ is the conjugate of α. This is called the ring of **quaternions** and generalizes the complex numbers in the following way: If $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\hat{i} = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$, $\hat{j} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, and $\hat{k} = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$ in $\mathbb{H}$, show that every

quaternion q has a unique representation in the form $q = aI + b\hat{i} + c\hat{j} + d\hat{k}$, where $a, b, c, d \in \mathbb{R}$. Show further that $\hat{i}^2 = \hat{j}^2 = \hat{k}^2 = \hat{i}\,\hat{j}\hat{k} = -I$ and that these relations determine the multiplication in $\mathbb{H}$. If $0 \neq q \in \mathbb{H}$, show that $q^{-1} = \dfrac{1}{a^2 + b^2 + c^2 + d^2} q^*$, where $q^* = aI - b\hat{i} - c\hat{j} - d\hat{k}$, so that $\mathbb{H}$ is a noncommutative skew field.

8.37. Find all the ring morphisms from $\mathbb{Z}$ to $\mathbb{Z}_6$.

8.38. Find all the ring morphisms from $\mathbb{Z}_{15}$ to $\mathbb{Z}_3$.

8.39. Find all the ring morphisms from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z} \times \mathbb{Z}$.

8.40. Find all the ring morphisms from $\mathbb{Z}_7$ to $\mathbb{Z}_4$.

8.41. If $(A, +)$ is an abelian group, the set of endomorphisms of A, $\mathrm{End}(A)$, consists of all the group morphisms from A to itself. Show that $(\mathrm{End}(A), +, \circ)$ is a ring under addition and composition, where $(f + g)(a) = f(a) + g(a)$, for $f, g \in \mathrm{End}(A)$. This is called the **endomorphism ring** of A.

8.42. Describe the endomorphism ring $\mathrm{End}(\mathbb{Z}_2 \times \mathbb{Z}_2)$. Is it commutative?

8.43. Prove that $10^n \equiv 1 \bmod 9$ for all $n \in \mathbb{N}$. Then prove that an integer is divisible by 9 if and only if the sum of its digits is divisible by 9.

8.44. Find the number of nonisomorphic rings with three elements.

8.45. Prove that $R[x] \cong R[y]$.

8.46. Prove that $R[x, y] \cong R[y, x]$.

8.47. Let $(R, +, \cdot)$ be a ring. Define the operations $\oplus$ and $\circ$ on R by

$$r \oplus s = r + s + 1 \quad \text{and} \quad r \circ s = r \cdot s + r + s.$$

(a) Prove that $(R, \oplus, \circ)$ is a ring.

(b) What are the additive and multiplicative identities of $(R, \oplus, \circ)$?

(c) Prove that $(R, \oplus, \circ)$ is isomorphic to $(R, +, \cdot)$.

8.48. Let a and b be elements of a commutative ring. For each positive integer n, prove the **binomial theorem**:

$$(a + b)^n = a^n + \binom{n}{1} a^{n-1}b + \cdots + \binom{n}{k} a^{n-k}b^k + \cdots + b^n.$$

8.49. Let $(R, +, \cdot)$ be an algebraic object that satisfies all the axioms for a ring except for the multiplicative identity. Define addition and multiplication in $R \times \mathbb{Z}$ by

$$(a, n) + (b, m) = (a + b, n + m) \quad \text{and}$$
$$(a, n) \cdot (b, m) = (ab + ma + nb, nm).$$

Show that $(R \times \mathbb{Z}, +, \cdot)$ is a ring that contains a subset in one-to-one correspondence with R that has all the properties of the algebraic object $(R, +, \cdot)$.

8.50. If R and S are commutative rings, prove that the ring of sequences $(R \times S)^{\mathbb{N}}$ is isomorphic to $R^{\mathbb{N}} \times S^{\mathbb{N}}$.

8.51. If F is a field, show that the field of fractions of F is isomorphic to F.

8.52. Describe the field of fractions of the ring $(\{a + ib | a, b \in \mathbb{Z}\}, +, \cdot)$.

8.53. Let $(S, *)$ be a commutative semigroup that satisfies the cancellation law; that is, $a * b = a * c$ implies that $b = c$. Show that $(S, *)$ can be embedded in a group.

8.54. Let $T = \{f : \mathbb{R} \to \mathbb{R} | f(x) = a \cos x + b \sin x, a, b \in \mathbb{R}\}$. Define addition of two such trigonometric functions in the usual way and define convolution by

$$(f * g)(x) = \int_0^{2\pi} f(t)g(x - t)\, dt.$$

Show that $(T, +, *)$ is a field.

8.55. Let $T_n = \left\{ f : \mathbb{R} \to \mathbb{R} | f(x) = \dfrac{a_0}{2} + \displaystyle\sum_{r=1}^{n} (a_r \cos rx + b_r \sin rx), a_r, b_r \in \mathbb{R} \right\}$. Show that $(T_n, +, *)$ is a commutative ring where addition and convolution are defined as in Exercise 8.54. What is the multiplicative identity? Is the ring an integral domain?

8.56. If R is any ring, define $R(i) = \{a + bi | a, b \in R\}$ to be the set of all formal sums $a + bi$, where a and b are in R. As in $\mathbb{C}$, we declare that $a + bi = a_1 + b_1 i$ if and only if $a = a_1$ and $b = b_1$. If we insist that $i^2 = -1$ and $ai = ia$ for all $a \in R$, then the ring axioms determine the addition and multiplication in $R(i)$:

$$(r + si) + (r_1 + s_1 i) = (r + r_1) + (s + s_1)i$$
$$(r + si)(r_1 + s_1 i) = (rr_1 - ss_1) + (rs_1 + sr_1)i.$$

Thus, for example, $\mathbb{R}(i) = \mathbb{C}$.

(a) Show that $R(i)$ is a ring, commutative if R is commutative.

(b) If R is commutative, show that $a + bi$ is has an inverse in $R(i)$ if and only if $a^2 + b^2$ has an inverse in R.

(c) Show that $\mathbb{Z}_3(i)$ is a field of nine elements.

(d) Is $\mathbb{C}(i)$ a field? Is $\mathbb{Z}_5(i)$ a field? Give reasons.

8.57. If R is a ring call $e \in R$ an **idempotent** if $e^2 = e$. Call R "tidy" if some positive power of every element is an idempotent.

(a) Show that every finite ring is tidy. [*Hint*: If $a \in R$, show that $a^{m+n} = a^m$ for some $n \geqslant 1$.]

(b) If R is tidy, show that $uv = 1$ in R implies that $vu = 1$.

(c) If R is a commutative tidy ring, show that every element of R is either invertible or a zero divisor.

9

POLYNOMIAL AND EUCLIDEAN RINGS

Polynomial functions and the solution of polynomial equations are a basic part of mathematics. One of the important uses of ring and field theory is to extend a field to a larger field so that a given polynomial has a root. For example, the complex number field can be obtained by enlarging the real field so that all quadratic equations will have solutions.

Before we are able to extend fields, we need to investigate the ring of polynomials, $F[x]$, with coefficients in a field F. This polynomial ring has many properties in common with the ring of integers; both $F[x]$ and $\mathbb{Z}$ are integral domains, but not fields. Moreover, both rings have division and euclidean algorithms. These algorithms are extremely useful, and rings with such algorithms are called *euclidean rings*.

EUCLIDEAN RINGS

Long division of integers gives a method for dividing one integer by another to obtain a quotient and a remainder. The fact that this is always possible is stated formally in the division algorithm.

Theorem 9.1. Division Algorithm for Integers. If a and b are integers and b is nonzero, then there exist unique integers q and r such that

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

Proof. If $b > 0$, then $|b| = b$, so this restates Theorem 7 in Appendix 1. If $b < 0$, then $-b > 0$, so the same theorem gives $a = q(-b) + r$, where $0 \leqslant r < (-b)$. Since $|b| = -b$ in this case, this gives $a = (-q)b + r$, where $0 \leqslant r < |b|$. □

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

The integer r is called the **remainder** in the division of a by b, and q is called the **quotient**.

What other rings, besides the integers, have a division algorithm? In a field, we can always divide any element exactly by a nonzero element. If a ring contains zero divisors, the cancellation property does not hold, and we cannot expect to obtain a unique quotient. This leaves integral domains, and the following kinds contain a useful generalization of the division algorithm.

An integral domain R is called a **euclidean ring** if for each nonzero element $a \in R$, there exists a nonnegative integer $\delta(a)$ such that:

(i) If a and b are nonzero elements of R, then $\delta(a) \leqslant \delta(ab)$.

(ii) For every pair of elements $a, b \in R$ with $b \neq 0$, there exist elements $q, r \in R$ such that

$$a = qb + r \quad \text{where} \quad r = 0 \quad \text{or} \quad \delta(r) < \delta(b). \qquad \text{(division algorithm)}$$

Theorem 9.1 shows that the ring $\mathbb{Z}$ of integers is a euclidean ring if we take $\delta(b) = |b|$, the absolute value of b, for all $b \in \mathbb{Z}$. A field is trivially a euclidean ring when $\delta(a) = 1$ for all nonzero elements a of the field. We now show that the ring of polynomials, with coefficients in a field, is a euclidean ring when we take $\delta(g(x))$ to be the degree of the polynomial $g(x)$.

Theorem 9.2. Division Algorithm for Polynomials. Let $f(x)$, $g(x)$ be elements of the polynomial ring $F[x]$, with coefficients in the *field* F. If $g(x)$ is not the zero polynomial, there exist unique polynomials $q(x), r(x) \in F[x]$ such that

$$f(x) = q(x) \cdot g(x) + r(x)$$

where either $r(x)$ is the zero polynomial or $\deg r(x) < \deg g(x)$.

Proof. If $f(x)$ is the zero polynomial or $\deg f(x) < \deg g(x)$, then writing $f(x) = 0 \cdot g(x) + f(x)$, we see that the requirements of the algorithm are fulfilled.

If $\deg f(x) = \deg g(x) = 0$, then $f(x)$ and $g(x)$ are nonzero constant polynomials a_0 and b_0, respectively. Now $f(x) = (a_0 b_0^{-1}) g(x)$, and the algorithm holds.

We prove the other cases by induction on the degree of $f(x)$. Suppose that, when we divide by a fixed polynomial $g(x)$, the division algorithm holds for polynomials of degree less than n. Let $f(x) = a_0 + \cdots + a_n x^n$ and $g(x) = b_0 + \cdots + b_m x^m$ where $a_n \neq 0, b_m \neq 0$. If $n < m$, we have already shown that the algorithm holds.

Suppose that $n \geqslant m$ and put

$$f_1(x) = f(x) - a_n b_m^{-1} x^{n-m} g(x)$$

so that deg $f_1(x) < n$. By the induction hypothesis

$$f_1(x) = q_1(x) \cdot g(x) + r(x) \quad \text{where either } r(x) = 0 \quad \text{or} \quad \deg r(x) < \deg g(x).$$

Hence $f(x) = a_n b_m^{-1} x^{n-m} g(x) + f_1(x) = \{a_n b_m^{-1} x^{n-m} + q_1(x)\} \cdot g(x) + r(x)$, which is a representation of the required form. The algorithm now follows by induction, starting with $n = m - 1$ if $m \neq 0$, or with $n = 0$ if $m = 0$.

The uniqueness of the quotient, $g(x)$, and of the remainder, $r(x)$, follows in a similar way to the uniqueness of the quotient and remainder in the division algorithm for integers (Theorem 7, Appendix 2). □

The quotient and remainder polynomials can be calculated by long division of polynomials.

Example 9.3. Divide $x^3 + 2x^2 + x + 2$ by $x^2 + 2$ in $\mathbb{Z}_3[x]$.

Solution. Write $\mathbb{Z}_3 = \{0, 1, 2\}$ for convenience.

$$\begin{array}{r|l} & x+2 \\ \hline x^2+2 & x^3+2x^2+\ x+2 \\ & \underline{x^3+\quad +2x} \\ & \quad 2x^2+2x+2 \\ & \quad \underline{2x^2\quad +1} \\ & \quad\quad 2x+1 \end{array}$$

Hence $x^3 + 2x^2 + x + 2 = (x + 2)(x^2 + 2) + (2x + 1)$. □

If we divide by a polynomial of degree 1, the remainder must be a constant. This constant can be found as follows.

Theorem 9.4. Remainder Theorem. The remainder when the polynomial $f(x)$ is divided by $(x - \alpha)$ in $F[x]$ is $f(\alpha)$.

Proof. By the division algorithm, there exist $q(x)$, $r(x) \in F[x]$ with $f(x) = q(x)(x - \alpha) + r(x)$, where $r(x) = 0$ or deg $r(x) < 1$. The remainder is therefore a constant $r_0 \in F$ and $f(x) = q(x)(x - \alpha) + r_0$. Substituting α for x, we obtain the result $f(\alpha) = r_0$. □

Theorem 9.5. Factor Theorem. The polynomial $(x - \alpha)$ is a factor of $f(x)$ in $F[x]$ if and only if $f(\alpha) = 0$.

Proof. We can write $f(x) = q(x)(x - \alpha)$ for some $q(x) \in F[x]$ if and only if $f(x)$ has remainder 0 when divided by $(x - \alpha)$. By the remainder theorem, this happens if and only if $f(\alpha) = 0$. □

An element α is called a **root** of a polynomial $f(x)$ if $f(\alpha) = 0$. The factor theorem shows that $(x - \alpha)$ is a factor of $f(x)$ if and only if α is a root of $f(x)$.

Theorem 9.6. A polynomial of degree n over a field F has at most n roots in F.

Proof. We prove the theorem by induction on the degree n. A polynomial of degree 0 consists of only a nonzero constant and therefore has no roots.

Assume that the theorem is true for polynomials of degree $n - 1$ and let $f(x) \in F[x]$ be a polynomial of degree n. If $f(x)$ has no roots, the theorem holds. If $f(x)$ does have roots, let α be one such root. By the factor theorem, we can write

$$f(x) = (x - \alpha)g(x),$$

and by Proposition 8.22, deg $g(x) = n - 1$.

Since the field F has no zero divisors, $f(\beta) = 0$ if and only if $(\beta - \alpha) = 0$ or $g(\beta) = 0$. Therefore, any root of $f(x)$ is either equal to α or is a root of $g(x)$. By the induction hypothesis, $g(x)$ has, at most, $n - 1$ roots, so $f(x)$ has, at most, n roots. □

Example 9.7. Show that the ring of gaussian integers, $\mathbb{Z}[i] = \{a + ib | a, b \in \mathbb{Z}\}$, is a euclidean ring with $\delta(a + ib) = a^2 + b^2$.

Solution. $\mathbb{Z}[i]$ is a subring of the complex numbers, $\mathbb{C}$, and therefore is an integral domain.

If $z \in \mathbb{Z}[i]$, then $\delta(z) = z\overline{z}$, where $\overline{z}$ is the conjugate of z in the complex numbers. For any nonzero complex number z, $\delta(z) > 0$, and for two nonzero gaussian integers z and w, $\delta(z \cdot w) = \delta(z) \cdot \delta(w)$.

To prove the division algorithm in $\mathbb{Z}[i]$, let z and w be gaussian integers where $w \neq 0$. Then z/w is a complex number, $c + id$, where $c, d \in \mathbb{Q}$. Choose integers, a, b as in Figure 9.1 so that $|c - a| \leqslant \frac{1}{2}$ and $|d - b| \leqslant \frac{1}{2}$. Then $z/w = a + ib + [(c - a) + i(d - b)]$ so $z = (a + ib)w + [(c - a) + i(d - b)]w$. Now $\delta([(c - a) + i(d - b)]w) =$

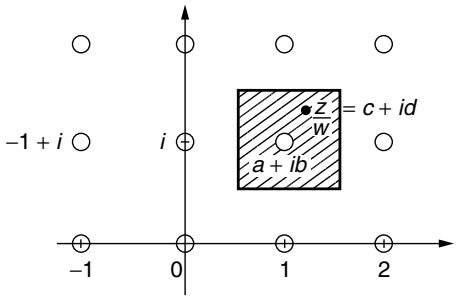


Figure 9.1. Complex numbers with the elements of $\mathbb{Z}[i]$ circled.

$\delta((c-a)+i(d-b))\delta(w)=[(c-a)^2+(d-b)^2]\delta(w) \leqslant (\frac{1}{4}+\frac{1}{4})\delta(w) < \delta(w)$.
Hence $\mathbb{Z}[i]$ is a euclidean ring. □

EUCLIDEAN ALGORITHM

The division algorithm allows us to generalize the concepts of divisors and greatest common divisors to any euclidean ring. Furthermore, we can produce a euclidean algorithm that will enable us to calculate greatest common divisors.

If a, b, q are three elements in an integral domain such that $a = qb$, we say that b **divides** a or that b is a **factor** of a and write $b|a$. For example, $(2+i)|(7+i)$ in the gaussian integers, $\mathbb{Z}[i]$, because $7+i = (3-i)(2+i)$.

Proposition 9.8. Let a, b, c be elements in an integral domain R.

(i) If $a|b$ and $a|c$, then $a|(b+c)$.
(ii) If $a|b$, then $a|br$ for any $r \in R$.
(iii) If $a|b$ and $b|c$, then $a|c$.

Proof. These results follow immediately from the definition of divisibility. □

By analogy with $\mathbb{Z}$, if a and b are elements in an integral domain R, then the element $g \in R$ is called a **greatest common divisor** of a and b, and is written $g = \gcd(a, b)$, if the following hold:

(i) $g|a$ and $g|b$.
(ii) If $c|a$ and $c|b$, then $c|g$.

The element $l \in R$ is called a **least common multiple** of a and b, and is written $l = \operatorname{lcm}(a, b)$, if the following hold:

(i) $a|l$ and $b|l$.
(ii) If $a|k$ and $b|k$, then $l|k$.

For example, 4 and -4 are greatest common divisors, and 60 and -60 are least common multiples, of 12 and 20 in $\mathbb{Z}$. Note that in $\mathbb{Z}$ it is customary to choose the positive value in each case to make it unique (see Appendix 2).

Theorem 9.9. Let R be a euclidean ring. Any two elements a and b in R have a greatest common divisor g. Moreover, there exist $s, t \in R$ such that

$$g = sa + tb.$$

Proof. If a and b are both zero, their greatest common divisor is zero, because $r|0$ for any $r \in R$.

Suppose that at least one of a and b is nonzero. By the well-ordering axiom (Appendix 2), let g be a nonzero element for which $\delta(g)$ is minimal in the set $I = \{xa + yb | x, y \in R\}$. We can write $g = sa + tb$ for some $s, t \in R$.

Since R is a euclidean ring, $a = hg + r$, where $r = 0$ or $\delta(r) < \delta(g)$. Therefore, $r = a - hg = a - h(sa + tb) = (1 - hs)a - htb \in I$. Since g was an element for which $\delta(g)$ was minimal in I, it follows that r must be zero, and $g|a$. Similarly, $g|b$.

If $c|a$ and $c|b$, so that $a = kc$ and $b = lc$, then $g = sa + tb = skc + tlc = (sk + tl)c$ and $c|g$. Therefore, $g = \gcd(a, b)$. □

Theorem 9.9 shows that greatest common divisors exist in any euclidean ring, but does not give a method for finding them. In fact, they can be computed using the following general euclidean algorithm.

Theorem 9.10. Euclidean Algorithm. Let a, b be elements of a euclidean ring R and let b be nonzero. By repeated use of the division algorithm, we can write

$$\begin{aligned}
a &= bq_1 + r_1 && \text{where} && \delta(r_1) < \delta(b) \\
b &= r_1q_2 + r_2 && \text{where} && \delta(r_2) < \delta(r_1) \\
r_1 &= r_2q_3 + r_3 && \text{where} && \delta(r_3) < \delta(r_2) \\
&\vdots && && \vdots \\
r_{k-2} &= r_{k-1}q_k + r_k && \text{where} && \delta(r_k) < \delta(r_{k-1}) \\
r_{k-1} &= r_kq_{k+1} + 0.
\end{aligned}$$

If $r_1 = 0$, then $b = \gcd(a, b)$; otherwise, $r_k = \gcd(a, b)$.

Furthermore, elements $s, t \in R$ such that

$$\gcd(a, b) = sa + tb$$

can be found by starting with the equation $r_k = r_{k-2} - r_{k-1}q_k$ and successively working up the sequence of equations above, each time replacing r_i in terms of r_{i-1} and r_{i-2}.

Proof. This algorithm must terminate, because $\delta(b), \delta(r_1), \delta(r_2), \ldots$ is a decreasing sequence of nonnegative integers; thus, $r_{k+1} = 0$ for some $k + 1$. The proof of the algorithm follows as in the proof of Theorem 10 in Appendix 2. □

Example 9.11. Find the greatest common divisor of 713 and 253 in $\mathbb{Z}$ and find two integers s and t such that

$$713s + 253t = \gcd(713, 253).$$

Solution. By the division algorithm, we have

$$\begin{aligned}
&\text{(i)}\ 713 = 2 \cdot 253 + 207 && a = 713, b = 253, r_1 = 207\\
&\text{(ii)}\ 253 = 1 \cdot 207 + 46 && r_2 = 46\\
&\text{(iii)}\ 207 = 4 \cdot 46 + 23 && r_3 = 23\\
&\qquad 46 = 2 \cdot 23 + 0. && r_4 = 0
\end{aligned}$$

The last nonzero remainder is the greatest common divisor. Hence $\gcd(713, 253) = 23$.

We can find the integers s and t by using equations (i)–(iii). We have

$$\begin{aligned}
23 &= 207 - 4 \cdot 46 && \text{from equation (iii)}\\
&= 207 - 4(253 - 207) && \text{from equation (ii)}\\
&= 5 \cdot 207 - 4 \cdot 253\\
&= 5 \cdot (713 - 2 \cdot 253) - 4 \cdot 253 && \text{from equation (i)}\\
&= 5 \cdot 713 - 14 \cdot 253.
\end{aligned}$$

Therefore, $s = 5$ and $t = -14$. □

Example 9.12. Find a greatest common divisor, $g(x)$, of $a(x) = 2x^4 + 2$ and $b(x) = x^5 + 2$ in $\mathbb{Z}_3[x]$, and find $s(x), t(x) \in \mathbb{Z}_3[x]$, so that

$$g(x) = s(x) \cdot (2x^4 + 2) + t(x) \cdot (x^5 + 2).$$

Solution. By repeated use of the division algorithm (see below), we have:

$$\begin{aligned}
&\text{(i)}\ x^5 + 2 = (2x)(2x^4 + 2) + (2x + 2).\\
&\text{(ii)}\ 2x^4 + 2 = (x^3 + 2x^2 + x + 2)(2x + 2) + 1.\\
&\text{(iii)}\ 2x + 2 = (2x + 2) \cdot 1 + 0.
\end{aligned}$$

Hence $\gcd(a(x), b(x)) = 1$. From equation (ii) we have

$$\begin{aligned}
1 &= 2x^4 + 2 - (x^3 + 2x^2 + x + 2)(2x + 2)\\
&= 2x^4 + 2 - (x^3 + 2x^2 + x + 2)\{x^5 + 2 - (2x)(2x^4 + 2)\}\\
&\qquad \text{from equation (i)}\\
&= (2x^4 + x^3 + 2x^2 + x + 1)(2x^4 + 2) + (2x^3 + x^2 + 2x + 1)(x^5 + 2).
\end{aligned}$$

Therefore,

$$s(x) = 2x^4 + x^3 + 2x^2 + x + 1$$

and

$$t(x) = 2x^3 + x^2 + 2x + 1.$$

$$
\begin{array}{r|l}
 & 2x \\
\cline{2-2}
2x^4+2 & x^5+0x^4+0x^3+0x^2+0x+2 \\
 & x^5 \qquad\qquad +\ x \\
\cline{2-2}
 & 2x+2
\end{array}
\qquad
\begin{array}{r|l}
 & x^3+2x^2+x+2 \\
\cline{2-2}
2x+2 & 2x^4+0x^3+0x^2+0x+2 \\
 & 2x^4+2x^3 \\
\cline{2-2}
 & x^3 \qquad\quad +2 \\
 & x^3+\ x^2 \\
\cline{2-2}
 & 2x^2 \quad +2 \\
 & 2x^2+2x \\
\cline{2-2}
 & x+2 \\
 & x+1 \\
\cline{2-2}
 & 1
\end{array}
$$

□

Example 9.13. Find a greatest common divisor of $a(x) = x^4 + x^3 + 3x - 9$ and $b(x) = 2x^3 - x^2 + 6x - 3$ in $\mathbb{Q}[x]$.

Solution. By the division algorithm we have (computation below)

$$a(x) = \left(\tfrac{1}{2}x + \tfrac{3}{4}\right) b(x) - \tfrac{9}{4}x^2 - \tfrac{27}{4}$$

and

$$b(x) = \left(-\tfrac{8}{9}x + \tfrac{4}{9}\right)\left(-\tfrac{9}{4}x^2 - \tfrac{27}{4}\right).$$

Hence

$$\gcd(a(x), b(x)) = -\tfrac{9}{4}x^2 - \tfrac{27}{4}.$$

$$
\begin{array}{r|l}
 & \frac{1}{2}x+\frac{3}{4} \\
\cline{2-2}
2x^3-x^2+6x-3 & x^4+\ x^3+0x^2+3x-\ 9 \\
 & x^4-\frac{1}{2}x^3+3x^2-\frac{3}{2}x \\
\cline{2-2}
 & \frac{3}{2}x^3-3x^2+\frac{9}{2}x-\ 9 \\
 & \frac{3}{2}x^3-\frac{3}{4}x^2+\frac{9}{2}x-\ \frac{9}{4} \\
\cline{2-2}
 & -\frac{9}{4}x^2 \qquad -\frac{27}{4}
\end{array}
\qquad
\begin{array}{r|l}
 & -\frac{8}{9}x+\frac{4}{9} \\
\cline{2-2}
-\frac{9}{4}x^2-\frac{27}{4} & 2x^3-x^2+6x-3 \\
 & 2x^3 \qquad +6x \\
\cline{2-2}
 & -x^2 \qquad -3 \\
 & -x^2 \qquad -3 \\
\cline{2-2}
 & 0
\end{array}
$$

□

UNIQUE FACTORIZATION

One important property of the integers, commonly known as the *fundamental theorem of arithmetic*, states that every integer greater than 1 can be written as

a finite product of prime numbers, and furthermore, this product is unique up to the ordering of the primes (see Theorem 13 in Appendix 2). In this section, we prove a similar result for any euclidean ring.

Let R be a commutative ring. An element u is called an **invertible element** (or **unit**) of R if there exists an element $v \in R$ such that $uv = 1$. The invertible elements in a ring R are those elements with multiplicative inverses in R. Denote the set of invertible elements of R by R^*. If R is a field, every nonzero element is invertible and $R^* = R - \{0\}$.

The invertible elements in the integers are ± 1. If F is a field, the invertible polynomials in $F[x]$ are the nonzero constant polynomials, that is, the polynomials of degree 0. The set of invertible elements in the gaussian integers is $\mathbb{Z}[i]^* = \{\pm 1, \pm i\}$.

Proposition 9.14. For any commutative ring R, the invertible elements form an abelian group, $(R^*, \cdot)$, under multiplication.

Proof. Let $u_1, u_2 \in R^*$ and let $u_1 v_1 = u_2 v_2 = 1$. Then $(u_1 u_2)(v_1 v_2) = 1$; thus $u_1 u_2 \in R^*$. The group axioms follow immediately. □

Two elements in a euclidean ring may have many greatest common divisors. For example, in $\mathbb{Q}[x]$, $x + 1$, $2x + 2$, and $\frac{1}{3}x + \frac{1}{3}$ are all greatest common divisors of $x^2 + 2x + 1$ and $x^2 - 1$. However, they can all be obtained from one another by multiplying by invertible elements.

Lemma 9.15. If $a|b$ and $b|a$ in an integral domain R, then $a = ub$, where u is an invertible element.

Proof. Since $a|b$, $b = va$ for $v \in R$ so if $a = 0$, then $b = 0$ and $a = b$. If $a \neq 0$, then $a = ub$ for $u \in R$ since $b|a$. Therefore, $a = ub = uva$; thus $a(uv - 1) = 0$. As $a \neq 0$ and R has no zero divisors, $uv = 1$ and u is invertible. □

Lemma 9.16. If g_2 is a greatest common divisor of a and b in the euclidean ring R, then g_1 is also a greatest common divisor of a and b if and only if $g_1 = ug_2$, where u is invertible.

Proof. If $g_1 = ug_2$ where $uv = 1$, then $g_2 = vg_1$. Hence $g_2|g_1$ and $g_1|g_2$ if and only if $g_1 = ug_2$. The result now follows from the definition of a greatest common divisor. □

Lemma 9.17. If a and b are elements in a euclidean ring R, then $\delta(a) = \delta(ab)$ if and only if b is invertible. Otherwise, $\delta(a) < \delta(ab)$.

Proof. If b is invertible and $bc = 1$, then $\delta(a) \leqslant \delta(ab) \leqslant \delta(abc) = \delta(a)$. Hence $\delta(a) = \delta(ab)$.

If b is not invertible, ab does not divide a and $a = qab + r$, where $\delta(r) < \delta(ab)$. Now $r = a(1 - qb)$; thus $\delta(a) \leqslant \delta(r)$. Therefore, $\delta(a) < \delta(ab)$. □

A noninvertible element p in a euclidean ring R is said to be **irreducible** if, whenever $p = ab$, either a or b is invertible in R. The irreducible elements in the integers are the prime numbers together with their negatives.

Lemma 9.18. Let R be a euclidean ring. If $a, b, c \in R$, $\gcd(a, b) = 1$ and $a|bc$, then $a|c$.

Proof. By Theorem 9.9, we can write $1 = sa + tb$, where $s, t \in R$. Therefore $c = sac + tbc$, so $a|c$ because $a|bc$. □

Proposition 9.19. If p is irreducible in the euclidean ring R and $p|ab$, then $p|a$ or $p|b$.

Proof. For any $a \in R$, write $d = \gcd(a, p)$. Then $d|p$, say $p = d \cdot h$. Since p is irreducible, either d or h is invertible, and so either $d = 1$ or p. Hence if p does not divide a, then $d = 1$, and it follows from Lemma 9.18 that $p|b$. □

Theorem 9.20. Unique Factorization Theorem. Every nonzero element in a euclidean ring R is either an invertible element or can be written as the product of a finite number of irreducibles. In such a product, the irreducibles are uniquely determined up to the order of the factors and up to multiplication by invertible elements.

Proof. We proceed by induction on $\delta(a)$ for $a \in R$. The least value of $\delta(a)$ for nonzero a is $\delta(1)$, because 1 divides any other element. Suppose that $\delta(a) = \delta(1)$. Then $\delta(1 \cdot a) = \delta(1)$ and, by Lemma 9.17, a is invertible.

By the induction hypothesis, suppose that all elements $x \in R$, with $\delta(x) < \delta(a)$, are either invertible or can be written as a product of irreducibles. We now prove this for the element a.

If a is irreducible, there is nothing to prove. If not, we can write $a = bc$, where neither b nor c is invertible. By Lemma 9.17, $\delta(b) < \delta(bc) = \delta(a)$ and $\delta(c) < \delta(bc) = \delta(a)$. By the induction hypothesis, b and c can each be written as a product of irreducibles, and hence a can also be written as a product of irreducibles.

To prove the uniqueness, suppose that

$$a = p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m,$$

where each p_i and q_j is irreducible. Now $p_1|a$ and so $p_1|q_1 q_2 \cdots q_m$. By an extension of Proposition 9.19 to m factors, p_1 divides some q_i. Rearrange the q_i, if necessary, so that $p_1|q_1$. Therefore, $q_1 = u_1 p_1$ where u_1 is invertible, because p_1 and q_1 are both irreducible.

Now $a = p_1 p_2 \cdots p_n = u_1 p_1 q_2 \cdots q_m$; thus $p_2 \cdots p_n = u_1 q_2 \cdots q_m$. Proceed inductively to show that $p_i = u_i q_i$ for all i, where each u_i is invertible.

If $m < n$, we would obtain the relation $p_{m+1} \cdots p_n = u_1 u_2 \cdots u_m$, which is impossible because irreducibles cannot divide an invertible element. If $m > n$, we would obtain

$$1 = u_1 u_2 \cdots u_n q_{n+1} \cdots q_m,$$

which is again impossible because an irreducible cannot divide 1. Hence $m = n$, and the primes $p_1, p_2, \ldots, p_n$ are the same as $q_1, q_2, \ldots, q_m$ up to a rearrangement and up to multiplication by invertible elements. □

When the euclidean ring is the integers, the theorem above yields the fundamental theorem of arithmetic referred to earlier. The ring of polynomials over a field and the gaussian integers also have this unique factorization property enjoyed by the integers. However, the integral domain

$$\mathbb{Z}[\sqrt{-3}] = \{a + b\sqrt{-3} \mid a, b \in \mathbb{Z}\},$$

which is a subring of $\mathbb{C}$, does not have the unique factorization property. For example,

$$4 = 2 \cdot 2 = (1 + \sqrt{-3}) \cdot (1 - \sqrt{-3}),$$

whereas 2, $1 + \sqrt{-3}$, and $1 - \sqrt{-3}$ are all irreducible. Therefore, $\mathbb{Z}[\sqrt{-3}]$ cannot be a euclidean ring.

FACTORING REAL AND COMPLEX POLYNOMIALS

The question of whether or not a polynomial is irreducible will be crucial in Chapter 10 when we extend number fields by adjoining roots of a polynomial. We therefore investigate different methods of factoring polynomials over various coefficient fields.

A polynomial $f(x)$ of positive degree is said to be **reducible over** the field F if it can be factored into two polynomials of positive degree in $F[x]$. If it cannot be so factored, $f(x)$ is called **irreducible over** F, and $f(x)$ is an irreducible element of the ring $F[x]$. It is important to note that reducibility depends on the field F. The polynomial $x^2 + 1$ is irreducible over $\mathbb{R}$ but reducible over $\mathbb{C}$.

The following basic theorem, first proved by Gauss in his doctoral thesis in 1799, enables us to determine which polynomials are irreducible in $\mathbb{C}[x]$ and in $\mathbb{R}[x]$.

Theorem 9.21. Fundamental Theorem of Algebra. If $f(x)$ is a polynomial in $\mathbb{C}[x]$ of positive degree, then $f(x)$ has a root in $\mathbb{C}$.

A proof of this theorem is given in Nicholson [11] using the fact from analysis that a cubic real polynomial has a real root.

The following useful theorem shows that the complex roots of *real* polynomials occur in conjugate pairs.

Theorem 9.22. (i) If $z = a + ib$ is a complex root of the real polynomial $f(x) \in \mathbb{R}[x]$, then its conjugate $\overline{z} = a - ib$ is also a root. Thus the real polynomial $(x - z)(x - \overline{z}) = x^2 - 2ax + (a^2 + b^2)$ is a factor of $f(x)$.

(ii) If $a, b, c \in \mathbb{Q}$ and $a + b\sqrt{c}$ is an irrational root of the rational polynomial $f(x) \in \mathbb{Q}[x]$, then $a - b\sqrt{c}$ is also a root, and the rational polynomial $x^2 - 2ax + (a^2 - b^2c)$ is a factor of $f(x)$.

Proof. (i) Let $g(x) = x^2 - 2ax + a^2 + b^2 = (x - z)(x - \overline{z})$. By the division algorithm in $\mathbb{R}[x]$, there exist real polynomials $q(x)$ and $r(x)$ such that

$$f(x) = q(x)g(x) + r(x) \quad \text{where} \quad r(x) = 0 \quad \text{or} \quad \deg r(x) < 2.$$

Hence $r(x) = r_0 + r_1x$ where $r_0, r_1 \in \mathbb{R}$. Now $z = a + ib$ is a root of $f(x)$ and of $g(x)$; therefore, it is also a root of $r(x)$, so $0 = r_0 + r_1(a + ib)$. Equating real and imaginary parts, we have $r_0 + r_1a = 0$ and $r_1b = 0$. But then

$$r(\overline{z}) = r(a - ib) = r_0 + r_1(a - ib) = r_0 + r_1a - ir_1b = 0.$$

Since $\overline{z}$ is a root of $r(x)$ and $g(x)$, it must be a root of $f(x)$.

If z is complex and not real, then $b \neq 0$. In this case $r_1 = 0$ and $r_0 = 0$; thus $g(x) | f(x)$.

(ii) This can be proved in a similar way to part (i). □

Theorem 9.23. (i) The irreducible polynomials in $\mathbb{C}[x]$ are the polynomials of degree 1.

(ii) The irreducible polynomials in $\mathbb{R}[x]$ are the polynomials of degree 1 together with the polynomials of degree 2 of the form $\mathrm{a}x^2 + \mathrm{b}x + \mathrm{c}$, where $b^2 < 4ac$.

Proof. (i) The polynomials of degree 0 are the invertible elements of $\mathbb{C}[x]$. By the fundamental theorem of algebra, any polynomial of positive degree has a root in $\mathbb{C}$ and hence a linear factor. Therefore, all polynomials of degree greater than 1 are reducible and those of degree 1 are the irreducibles.

(ii) The polynomials of degree 0 are the invertible elements of $\mathbb{R}[x]$. By part (i) and the unique factorization theorem, every real polynomial of positive degree can be factored into linear factors in $\mathbb{C}[x]$. By Theorem 9.22 (i), its nonreal roots fall into conjugate pairs, whose corresponding factors combine to give a quadratic factor in $\mathbb{R}[x]$ of the form $ax^2 + bx + c$, where $b^2 < 4ac$. Hence any real polynomial can be factored into real linear factors and real quadratic factors of the form above. □

Example 9.24. Find the kernel and image of the ring morphism $\psi: \mathbb{Q}[x] \to \mathbb{R}$ defined by $\psi(f(x)) = f(\sqrt{2})$.

Solution. If $p(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Q}[x]$, then

$$\begin{aligned}\psi(p(x)) &= a_0 + a_1\sqrt{2} + \cdots + a_n(\sqrt{2})^n \\ &= (a_0 + 2a_2 + 4a_4 + \cdots) + \sqrt{2}\,(a_1 + 2a_3 + 4a_5 + \cdots),\end{aligned}$$

so $\psi(p(x)) \in \mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} | a, b \in \mathbb{Q}\}$, where $\mathbb{Q}(\sqrt{2})$ is the subring of $\mathbb{R}$ defined in Example 8.3. Hence $\operatorname{Im}\psi \subseteq \mathbb{Q}(\sqrt{2})$, and $\operatorname{Im}\psi = \mathbb{Q}(\sqrt{2})$ because $\psi(a + bx) = a + b\sqrt{2}$.

If $p(x) \in \operatorname{Ker}\psi$, then $p(\sqrt{2}) = 0$; therefore, by Theorem 9.22(ii), $p(-\sqrt{2}) = 0$, and $p(x)$ contains a factor $(x^2 - 2)$. Conversely, if $p(x)$ contains a factor $(x^2 - 2)$, then $p(\sqrt{2}) = 0$ and $p(x) \in \operatorname{Ker}\psi$. Hence $\operatorname{Ker}\psi = \{(x^2 - 2)q(x) | q(x) \in \mathbb{Q}[x]\}$, that is, the set of all polynomials in $\mathbb{Q}[x]$ with $(x^2 - 2)$ as a factor. □

FACTORING RATIONAL AND INTEGRAL POLYNOMIALS

A rational polynomial can always be reduced to an integer polynomial by multiplying it by the least common multiple of the denominators of its coefficients. We now give various methods for determining whether an integer polynomial has rational roots or is irreducible over $\mathbb{Q}$.

Theorem 9.25. Rational Roots Theorem. Let $p(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}[x]$. If r/s is a rational root of $p(x)$ and $\gcd(r, s) = 1$, then:

(i) $r | a_0$.
(ii) $s | a_n$.

Proof. If $p(r/s) = 0$, then $a_0 + a_1(r/s) + \cdots + a_{n-1}(r/s)^{n-1} + a_n(r/s)^n = 0$, whence $a_0s^n + a_1rs^{n-1} + \cdots + a_{n-1}r^{n-1}s + a_nr^n = 0$. Therefore, $a_0s^n = -r(a_1s^{n-1} + \cdots + a_{n-1}r^{n-2}s + a_nr^{n-1})$; thus $r | a_0s^n$. Since $\gcd(r, s) = 1$, it follows from Lemma 9.18 that $r | a_0$. Similarly, $s | a_n$. □

Example 9.26. Factor $p(x) = 2x^3 + 3x^2 - 1$ in $\mathbb{Q}[x]$.

Solution. If $p(r/s) = 0$, then, by Theorem 9.25, $r | (-1)$ and $s | 2$. Hence $r = \pm 1$ and $s = \pm 1$ or ± 2, and the only possible values of r/s are $\pm 1, \pm 1/2$. Instead of testing all these values, we sketch the graph of $p(x)$ to find approximate roots. Differentiating, we have $p'(x) = 6x^2 + 6x = 6x(x + 1)$, so $p(x)$ has turning values at 0 and -1.

We see from the graph in Figure 9.2 that -1 is a double root and that there is one more positive root. If it is rational, it can only be $\frac{1}{2}$. Checking this in Table 9.1, we see that $\frac{1}{2}$ is a root; hence $p(x)$ factors as $(x + 1)^2(2x - 1)$. □

Example 9.27. Prove that $\sqrt[5]{2}$ is irrational.

TABLE 9.1

x	-1	0	1/2	1	2
$p(x)$	0	-1	0	4	27

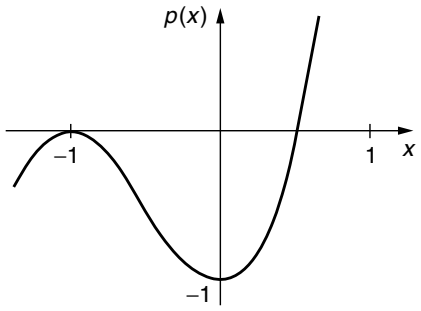


Figure 9.2. Graph of $p(x) = 2x^3 + 3x^2 - 1$.

TABLE 9.2

x	-2	-1	1	2
$x^5 - 2$	-34	-3	-1	30

Solution. Observe that $\sqrt[5]{2}$ is a root of $x^5 - 2$. If this polynomial has a rational root r/s, in its lowest terms, it follows from Theorem 9.25 that $r|(-2)$ and $s|1$. Hence the only possible rational roots are $\pm 1, \pm 2$. We see from Table 9.2 that none of these are roots, so all the roots of the polynomial $x^5 - 2$ must be irrational. □

Theorem 9.28. Gauss' Lemma. Let $P(x) = a_0 + \cdots + a_n x^n \in \mathbb{Z}[x]$. If $P(x)$ can be factored in $\mathbb{Q}[x]$ as $P(x) = q(x)r(x)$ with $q(x), r(x) \in \mathbb{Q}[x]$, then $P(x)$ can also be factored in $\mathbb{Z}[x]$.

Proof. Express the rational coefficients of $q(x)$ in their lowest terms and let u be the least common multiple of their denominators. Then $q(x) = (1/u)\overline{Q}(x)$, where $\overline{Q}(x) \in \mathbb{Z}[x]$. Let s be the greatest common divisor of all the coefficients of $\overline{Q}(x)$; write $q(x) = (s/u)Q(x)$, where $Q(x) \in \mathbb{Z}[x]$, and the greatest common divisor of its coefficients is 1. Write $r(x) = (t/v)R(x)$ in a similar way.

Now $P(x) = q(x)r(x) = \dfrac{s}{u}Q(x)\dfrac{t}{v}R(x) = \dfrac{st}{uv}Q(x)R(x)$, so $uvP(x) = stQ(x)R(x)$. To prove the theorem, we show that $uv|st$ by proving that no prime p in uv can divide all the coefficients of $Q(x)R(x)$.

Let $Q(x) = b_0 + \cdots + b_k x^k$ and $R(x) = c_0 + \cdots + c_l x^l$. Choose a prime p and let b_i and c_j be the first coefficients of $Q(x)$ and $R(x)$, respectively, that p fails to divide. These exist because $\gcd(b_0, \ldots, b_k) = 1$ and $\gcd(c_0, \ldots, c_l) = 1$. The coefficient of x^{i+j} in $Q(x)R(x)$ is

$$b_{i+j}c_0 + b_{i+j-1}c_1 + \cdots + b_{i+1}c_{j-1} + b_i c_j + b_{i-1}c_{j+1} + \cdots + b_0 c_{i+j}.$$

Now $p|c_0$, $p|c_1, \ldots, p|c_{j-1}$, $p|b_{i-1}$, $p|b_{i-2}, \ldots, p|b_0$ but $p \not| b_i c_j$ so this coefficient is not divisible by p. Hence the greatest common divisor of the coefficients of $Q(x)R(x)$ is 1; therefore, $uv|st$ and $P(x)$ can be factored in $\mathbb{Z}[x]$. □

Example 9.29. Factor $p(x) = x^4 - 3x^2 + 2x + 1$ into irreducible factors in $\mathbb{Q}[x]$.

Solution. By Theorem 9.25, the only possible rational roots are ± 1. However, these are not roots, so $p(x)$ has no linear factors.

Therefore, if it does factor, it must factor into two quadratics, and by Gauss' lemma these factors can be chosen to have integral coefficients. Suppose that

$$\begin{aligned} x^4 - 3x^2 + 2x + 1 &= (x^2 + ax + b)(x^2 + cx + d) \\ &= x^4 + (a + c)x^3 + (b + d + ac)x^2 + (bc + ad)x + bd. \end{aligned}$$

Thus we have to solve the following system for integer solutions:

$$a + c = 0, \quad b + d + ac = -3, \quad bc + ad = 2, \quad \text{and} \quad bd = 1.$$

Therefore, $b = d = \pm 1$ and $b(a + c) = 2$. Hence $a + c = \pm 2$, which is a contradiction. The polynomial cannot be factored into two quadratics and therefore is irreducible in $\mathbb{Q}[x]$. □

Theorem 9.30. Eisenstein's Criterion. Let $f(x) = a_0 + a_1 x + \cdots + a_n x^n \in \mathbb{Z}[x]$. Suppose that the following conditions all hold for some prime p:

(i) $p|a_0$, $p|a_1, \ldots, p|a_{n-1}$.
(ii) $p \not| a_n$.
(iii) $p^2 \not| a_0$.

Then $f(x)$ is irreducible over $\mathbb{Q}$.

Proof. Suppose that $f(x)$ is reducible. By Gauss' lemma, it factors as two polynomials in $\mathbb{Z}[x]$; that is,

$$f(x) = (b_0 + \cdots + b_r x^r)(c_0 + \cdots + c_s x^s),$$

where $b_i, c_j \in \mathbb{Z}$, $s > 0$, and $r + s = n$. Comparing coefficients, we see that $a_0 = b_0 c_0$. Now $p|a_0$, but $p^2 \not| a_0$, so p must divide b_0 or c_0 but not both. Without loss of generality, suppose that $p|b_0$ and $p \not| c_0$. Now p cannot divide all of $b_0, b_1, \ldots, b_r$, for then p would divide a_n. Let t be the smallest integer for which $p \not| b_t$; thus $1 \leqslant t \leqslant r < n$. Then $a_t = b_t c_0 + b_{t-1} c_1 + \cdots + b_1 c_{t-1} + b_0 c_t$ and $p|a_t$, $p|b_0$, $p|b_1, \ldots, p|b_{t-1}$. Hence $p|b_t c_0$. However, $p \not| b_t$ and $p \not| c_0$, so we have a contradiction, and the theorem is proved. □

For example, Eisenstein's criterion can be used to show that $x^5 - 2$, $x^7 + 2x^3 + 12x^2 - 2$ and $2x^3 + 9x - 3$ are all irreducible over $\mathbb{Q}$.

Example 9.31. Show that $\phi(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$ is irreducible over $\mathbb{Q}$ for any prime p. This is called a **cyclotomic polynomial** and can be written $\phi(x) = (x^p - 1)/(x - 1)$.

Solution. We cannot apply Eisenstein's criterion to $\phi(x)$ as it stands. However, if we put $x = y + 1$, we obtain

$$\begin{aligned}\phi(y+1) &= \frac{1}{y}[(y+1)^p - 1] \\ &= y^{p-1} + \binom{p}{p-1} y^{p-2} + \binom{p}{p-2} y^{p-3} + \cdots + \binom{p}{2} y + p\end{aligned}$$

where $\binom{p}{k} = \dfrac{p!}{k!(p-k)!}$ is the binomial coefficient. Hence p divides $k!(p-k)!\binom{p}{k}$. If $1 \leqslant k \leqslant p-1$, the prime p does not divide $k!(p-k)!$, so it must divide $\binom{p}{k}$. Hence $\phi(y+1)$ is irreducible by Eisenstein's criterion, so $\phi(x)$ is irreducible. □

FACTORING POLYNOMIALS OVER FINITE FIELDS

The roots of a polynomial in $\mathbb{Z}_p[x]$ can be found by trying all the p possible values.

Example 9.32. Does $x^4 + 4 \in \mathbb{Z}_7[x]$ have any roots in $\mathbb{Z}_7$?

Solution. We see from Table 9.3 that $x^4 + 4$ is never zero and therefore has no roots in $\mathbb{Z}_7$. □

Proposition 9.33. A polynomial in $\mathbb{Z}_2[x]$ has a factor $x + 1$ if and only if it has an even number of nonzero coefficients.

Proof. Let $p(x) = a_0 + a_1x + \cdots + a_nx^n \in \mathbb{Z}_2[x]$. By the factor theorem, $(x + 1)$ is a factor of $p(x)$ if and only if $p(1) = 0$. (Remember that $x - 1 = x + 1$

TABLE 9.3. Values Modulo 7

x	0	1	2	3	4	5	6
x^4	0	1	2	4	4	2	1
$x^4 + 4$	4	5	6	1	1	6	5

in $\mathbb{Z}_2[x]$.) Now $p(1) = a_0 + a_1 + \cdots + a_n$, which is zero in $\mathbb{Z}_2$ if and only if $p(x)$ has an even number of nonzero coefficients. □

Example 9.34. Find all the irreducible polynomials of degree less than or equal to 4 over $\mathbb{Z}_2$.

Solution. Degree 1 polynomials are irreducible; in $\mathbb{Z}_2[x]$ we have x and $x + 1$.

Let $p(x) = a_0 + \cdots + a_n x^n \in \mathbb{Z}_2[x]$. If $p(x)$ has degree n, then a_n is nonzero, so $a_n = 1$. The only possible roots are 0 and 1. The element 0 is a root if and only if $a_0 = 0$, and 1 is a root if and only if $p(x)$ has an even number of nonzero terms. Hence the following are the polynomials of degrees 2, 3, and 4 in $\mathbb{Z}_2[x]$ with no linear factors:

$$\begin{array}{ll} x^2 + x + 1 & \text{(degree 2)} \\ x^3 + x + 1,\ x^3 + x^2 + 1 & \text{(degree 3)} \\ x^4 + x + 1,\ x^4 + x^2 + 1,\ x^4 + x^3 + 1,\ x^4 + x^3 + x^2 + x + 1 & \text{(degree 4).} \end{array}$$

If a polynomial of degree 2 or 3 is reducible, it must have a linear factor; hence the polynomials of degree 2 and 3 above are irreducible. If a polynomial of degree 4 is reducible, it either has a linear factor or is the product of two irreducible quadratic factors. Now there is only one irreducible quadratic in $\mathbb{Z}_2[x]$, and its square $(x^2 + x + 1)^2 = x^4 + x^2 + 1$ is reducible.

Hence the irreducible polynomials of degree $\leqslant 4$ over $\mathbb{Z}_2$ are x, $x + 1$, $x^2 + x + 1$, $x^3 + x + 1$, $x^3 + x^2 + 1$, $x^4 + x + 1$, $x^4 + x^3 + 1$, and $x^4 + x^3 + x^2 + x + 1$. □

For example, the polynomials of degree 4 in $\mathbb{Z}_2[x]$ factorize into irreducible factors as follows.

$$\begin{array}{ll} x^4 & = x^4 \\ x^4 + 1 & = (x + 1)^4 \\ x^4 + x & = x(x + 1)(x^2 + x + 1) \\ x^4 + x + 1 & \text{is irreducible} \\ x^4 + x^2 & = x^2(x + 1)^2 \\ x^4 + x^2 + 1 & = (x^2 + x + 1)^2 \\ x^4 + x^2 + x & = x(x^3 + x + 1) \\ x^4 + x^2 + x + 1 & = (x + 1)(x^3 + x^2 + 1) \\ x^4 + x^3 & = x^3(x + 1) \\ x^4 + x^3 + 1 & \text{is irreducible} \\ x^4 + x^3 + x & = x(x^3 + x^2 + 1) \\ x^4 + x^3 + x + 1 & = (x + 1)^2(x^2 + x + 1) \\ x^4 + x^3 + x^2 & = x^2(x^2 + x + 1) \\ x^4 + x^3 + x^2 + 1 & = (x + 1)(x^3 + x + 1) \\ x^4 + x^3 + x^2 + x & = x(x + 1)^3 \\ x^4 + x^3 + x^2 + x + 1 & \text{is irreducible} \end{array}$$

LINEAR CONGRUENCES AND THE CHINESE REMAINDER THEOREM

The euclidean algorithm for integers can be used to solve linear congruences. We first find the conditions for a single congruence to have a solution and then show how to find all its solutions, if they exist. We then present the Chinese remainder theorem, which gives conditions under which many simultaneous congruences, with coprime moduli, have solutions. These solutions can again be found by using the euclidean algorithm.

First let us consider a linear congruence of the form

$$ax \equiv b \bmod n.$$

This has a solution if and only if the equation

$$ax + ny = b$$

has integer solutions for x and y. The congruence is also equivalent to the equation $[a][x] = [b]$ in $\mathbb{Z}_n$.

Theorem 9.35. The equation $ax + ny = b$ has solutions for $x, y \in \mathbb{Z}$ if and only if $\gcd(a, n)|b$.

Proof. Write $d = \gcd(a, n)$. If $ax + ny = b$ has a solution, then $d|b$ because $d|a$ and $d|n$. Conversely, let $d|b$, say $b = k \cdot d$. By Theorem 9.9, there exist $s, t \in \mathbb{Z}$ such that $as + nt = d$. Hence $ask + ntk = k \cdot d$ and $x = sk$, $y = tk$ is a solution to $ax + ny = b$. □

The euclidean algorithm gives a practical way to find the integers s and t in Theorem 9.35. These can then be used to find *one* solution to the equation.

Theorem 9.36. The congruence $ax \equiv b \bmod n$ has a solution if and only if $d|b$, where $d = \gcd(a, n)$. Moreover, if this congruence does have at least one solution, the number of noncongruent solutions modulo n is d; that is, if $[a][x] = [b]$ has a solution in $\mathbb{Z}_n$, then it has d different solutions in $\mathbb{Z}_n$.

Proof. The condition for the existence of a solution follows immediately from Theorem 9.35. Now suppose that x_0 is a solution, so that $ax_0 \equiv b \bmod n$. Let $d = \gcd(a, n)$ and $a = da'$, $n = dn'$. Then $gcd(a', n') = 1$, so the following statements are all equivalent.

(i) x is a solution to the congruence $ax \equiv b \bmod n$.
(ii) x is a solution to the congruence $a(x - x_0) \equiv 0 \bmod n$.
(iii) $n|a(x - x_0)$.
(iv) $n'|a'(x - x_0)$.

(v) $n' | (x - x_0)$.
(vi) $x = x_0 + kn'$ for some $k \in \mathbb{Z}$.

Now $x_0, x_0 + n', x_0 + 2n', \ldots, x_0 + (d-1)n'$ form a complete set of noncongruent solutions modulo n, and there are d such solutions. □

Example 9.37. Find the inverse of [49] in the field $\mathbb{Z}_{53}$.

Solution. Let $[x] = [49]^{-1}$ in $\mathbb{Z}_{53}$. Then $[49] \cdot [x] = [1]$; that is, $49x \equiv 1 \mod 53$. We can solve this congruence by solving the equation $49x - 1 = 53y$, where $y \in \mathbb{Z}$. By using the euclidean algorithm we have

$$53 = 1 \cdot 49 + 4 \quad \text{and} \quad 49 = 12 \cdot 4 + 1.$$

Hence $\gcd(49, 53) = 1 = 49 - 12 \cdot 4 = 49 - 12(53 - 49) = 13 \cdot 49 - 12 \cdot 53$. Therefore, $13 \cdot 49 \equiv 1 \mod 53$ and $[49]^{-1} = [13]$ in $\mathbb{Z}_{53}$. □

Theorem 9.38. Chinese Remainder Theorem. Let $m = m_1 m_2 \cdots m_r$, where $\gcd(m_i, m_j) = 1$ if $i \neq j$. Then the system of simultaneous congruences

$$x \equiv a_1 \mod m_1, \quad x \equiv a_2 \mod m_2, \quad \ldots, \quad x \equiv a_r \mod m_r$$

always has an integral solution. Moreover, if b is one solution, the complete solution is the set of integers satisfying $x \equiv b \mod m$.

Proof. This result follows from the ring isomorphism

$$f: \mathbb{Z}_m \to \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2} \times \cdots \times \mathbb{Z}_{m_r}$$

of Theorem 8.20 defined by $f([x]_m) = ([x]_{m_1}, [x]_{m_2}, \ldots, [x]_{m_r})$. The integer x is a solution of the simultaneous congruences if and only if $f([x]_m) = ([a_1]_{m_1}, [a_2]_{m_2}, \ldots, [a_r]_{m_r})$. Therefore, there is always a solution, and the solution set consists of exactly one congruence class modulo m. □

One method of finding the solution to a set of simultaneous congruences is to use the euclidean algorithm repeatedly.

Example 9.39. Solve the simultaneous congruences $\begin{cases} x \equiv 36 \mod 41 \\ x \equiv 5 \mod 17 \end{cases}$.

Proof. Any solution to the first congruence is of the form $x = 36 + 41t$ where $t \in \mathbb{Z}$. Substituting this into the second congruence, we obtain

$$36 + 41t \equiv 5 \mod 17 \quad \text{that is,} \quad 41t \equiv -31 \mod 17.$$

Reducing modulo 17, we have $7t \equiv 3 \bmod 17$. Solving this by the euclidean algorithm, we have

$$17 = 2 \cdot 7 + 3 \quad \text{and} \quad 7 = 2 \cdot 3 + 1.$$

Therefore, $1 = 7 - 2(17 - 2 \cdot 7) = 7 \cdot 5 - 17 \cdot 2$ and $7 \cdot 5 \equiv 1 \bmod 17$. Hence $7 \cdot 15 \equiv 3 \bmod 17$, so $t \equiv 15 \bmod 17$ is the solution to $7t \equiv 3 \bmod 17$.

We have shown that if $x = 36 + 41t$ is a solution to both congruences, then $t = 15 + 17u$, where $u \in \mathbb{Z}$. That is,

$$x = 36 + 41t = 36 + 41(15 + 17u) = 651 + 697u$$

or $x \equiv 651 \bmod 697$ is the complete solution. □

Example 9.40. Find the smallest positive integer that has remainders 4, 3, and 1 when divided by 5, 7, and 9, respectively.

Solution. We have to solve the three simultaneous congruences

$$x \equiv 4 \bmod 5, \quad x \equiv 3 \bmod 7, \quad \text{and} \quad x \equiv 1 \bmod 9.$$

The first congruence implies that $x = 4 + 5t$, where $t \in \mathbb{Z}$. Substituting into the second congruence, we have

$$4 + 5t \equiv 3 \bmod 7.$$

Hence $5t \equiv -1 \bmod 7$. Now $5^{-1} = 3$ in $\mathbb{Z}_7$, so $t \equiv 3 \cdot (-1) \equiv 4 \bmod 7$. Therefore, $t = 4 + 7u$, where $u \in \mathbb{Z}$, and any integer satisfying the first two congruences is of the form

$$x = 4 + 5t = 4 + 5(4 + 7u) = 24 + 35u.$$

Substituting this into the third congruence, we have $24 + 35u \equiv 1 \bmod 9$ and $-u \equiv -23 \bmod 9$. Thus $u \equiv 5 \bmod 9$ and $u = 5 + 9v$ for some $v \in \mathbb{Z}$.

Hence any solution of the three congruences is of the form

$$x = 24 + 35u = 24 + 35(5 + 9v) = 199 + 315v.$$

The smallest positive solution is $x = 199$. □

The Chinese remainder theorem was known to ancient Chinese astronomers, who used it to date events from observations of various periodic astronomical phenomena. It is used in this computer age as a tool for finding integer solutions to integer equations and for speeding up arithmetic operations in a computer.

Addition of two numbers in conventional representation has to be carried out sequentially on the digits in each position; the digits in the ith position have to

be added before the digit to be carried over to the $(i + 1)$st position is known. One method of speeding up addition on a computer is to perform addition using residue representation, since this avoids delays due to carry digits.

Let $m = m_1 m_2 \cdots m_r$, where the integers m_i are coprime in pairs. The **residue representation** or **modular representation** of any number x in $\mathbb{Z}_m$ is the r-tuple $(a_1, a_2, \ldots, a_r)$, where $x \equiv a_i \bmod m_i$.

For example, every integer from 0 to 29 can be uniquely represented by its residues modulo 2, 3, and 5 in Table 9.4.

This residue representation corresponds exactly to the isomorphism

$$\mathbb{Z}_{30} \to \mathbb{Z}_2 \times \mathbb{Z}_3 \times \mathbb{Z}_5.$$

Since this is a ring isomorphism, addition and multiplication are performed simply by adding and multiplying each residue separately.

For example, to add 4 and 7 using residue representation, we have

$$(0, 1, 4) + (1, 1, 2) = (0 + 1, 1 + 1, 4 + 2) = (1, 2, 1).$$

Similarly, multiplying 4 and 7, we have

$$(0, 1, 4) \cdot (1, 1, 2) = (0 \cdot 1, 1 \cdot 1, 4 \cdot 2) = (0, 1, 3).$$

Fast adders can be designed using residue representation, because all the residues can be added simultaneously. Numbers can be converted easily into residue form; however, the reverse procedure of finding a number with a given residue representation requires the Chinese remainder theorem. See Knuth [19, Sec. 4.3.2] for further discussion of the use of residue representations in computers.

TABLE 9.4. Residue Representation of the Integers from 0 to 29

	Residues Modulo:				Residues Modulo:				Residues Modulo:		
x	2	3	5	x	2	3	5	x	2	3	5
0	0	0	0	10	0	1	0	20	0	2	0
1	1	1	1	11	1	2	1	21	1	0	1
2	0	2	2	12	0	0	2	22	0	1	2
3	1	0	3	13	1	1	3	23	1	2	3
4	0	1	4	14	0	2	4	24	0	0	4
5	1	2	0	15	1	0	0	25	1	1	0
6	0	0	1	16	0	1	1	26	0	2	1
7	1	1	2	17	1	2	2	27	1	0	2
8	0	2	3	18	0	0	3	28	0	1	3
9	1	0	4	19	1	1	4	29	1	2	4

EXERCISES

For Exercises 9.1 to 9.6 calculate the quotients and remainders.

9.1. Divide $3x^4 + 4x^3 - x^2 + 5x - 1$ by $2x^2 + x + 1$ in $\mathbb{Q}[x]$.
9.2. Divide $x^6 + x^4 - 4x^3 + 5x$ by $x^3 + 2x^2 + 1$ in $\mathbb{R}[x]$.
9.3. Divide $x^7 + x^6 + x^4 + x + 1$ by $x^3 + x + 1$ in $\mathbb{Z}_2[x]$.
9.4. Divide $2x^5 + x^4 + 2x^3 + x^2 + 2$ by $x^3 + 2x + 2$ in $\mathbb{Z}_3[x]$.
9.5. Divide $17 + 11i$ by $3 + 4i$ in $\mathbb{Z}[i]$.
9.6. Divide $20 + 8i$ by $7 - 2i$ in $\mathbb{Z}[i]$.

For Exercises 9.7 to 9.13, find the greatest common divisors of the elements a, b in the given euclidean ring, and find elements s, t in the ring so that $as + bt = gcd(a, b)$.

9.7. $a = 33$, $b = 42$ in $\mathbb{Z}$.
9.8. $a = 2891$, $b = 1589$ in $\mathbb{Z}$.
9.9. $a = 2x^3 - 4x^2 - 8x + 1$, $b = 2x^3 - 5x^2 - 5x + 2 \in \mathbb{Q}[x]$.
9.10. $a = x^6 - x^3 - 16x^2 + 12x - 2$, $b = x^5 - 2x^2 - 16x + 8 \in \mathbb{Q}[x]$.
9.11. $a = x^4 + x + 1$, $b = x^3 + x^2 + x \in \mathbb{Z}_3[x]$.
9.12. $a = x^4 + 2$, $b = x^3 + 3 \in \mathbb{Z}_5[x]$.
9.13. $a = 4 - i$, $b = 1 + i \in \mathbb{Z}[i]$.

For Exercises 9.14 to 9.17, find one solution to each equation with $x, y \in \mathbb{Z}$.

9.14. $15x + 36y = 3$.
9.15. $24x + 29y = 1$.
9.16. $24x + 29y = 6$.
9.17. $11x + 31y = 1$.

For Exercises 9.18 to 9.21, find the inverse to the element in the given field.

9.18. $[4]$ in $\mathbb{Z}_7$.
9.19. $[24]$ in $\mathbb{Z}_{29}$.
9.20. $[35]$ in $\mathbb{Z}_{101}$.
9.21. $[11]$ in $\mathbb{Z}_{31}$.

Find all integral solutions to the equations in Exercises 9.22 to 9.24.

9.22. $27x + 15y = 13$.
9.23. $12x + 20y = 14$.
9.24. $28x + 20y = 16$.

Factor the polynomials in Exercises 9.25 to 9.36 into irreducible factors in the given ring.

9.25. $x^5 - 1$ in $\mathbb{Q}[x]$.
9.26. $x^5 + 1$ in $\mathbb{Z}_2[x]$.
9.27. $x^4 + 1$ in $\mathbb{Z}_5[x]$.
9.28. $2x^3 + x^2 + 4x + 2$ in $\mathbb{Q}[x]$.
9.29. $x^4 - 9x + 3$ in $\mathbb{Q}[x]$.
9.30. $2x^3 + x^2 + 4x + 2$ in $\mathbb{C}[x]$.
9.31. $x^3 - 4x + 1$ in $\mathbb{Q}[x]$.
9.32. $x^4 + 3x^3 + 9x - 9$ in $\mathbb{Q}[x]$.
9.33. $x^8 - 16$ in $\mathbb{C}[x]$.
9.34. $x^8 - 16 \in \mathbb{R}[x]$.
9.35. $x^8 - 16$ in $\mathbb{Q}[x]$.
9.36. $x^8 - 16 \in \mathbb{Z}_{17}[x]$.
9.37. Find all irreducible polynomials of degree 5 over $\mathbb{Z}_2$.

9.38. Find an irreducible polynomial of degree 2 over $\mathbb{Z}_5$.

9.39. Find an irreducible polynomial of degree 3 over $\mathbb{Z}_7$.

9.40. Find the kernel and image of the ring morphism $\psi: \mathbb{R}[x] \to \mathbb{C}$ defined by $\psi(p(x)) = p(i)$, where $i = \sqrt{-1}$.

9.41. Find the kernel and image of the ring morphism $\psi: \mathbb{R}[x] \to \mathbb{C}$ defined by $\psi(p(x)) = p(1 + \sqrt{3}i)$.

In Exercises 9.42 to 9.47, are the polynomials irreducible in the given ring? Give reasons.

9.42. $x^3 + x^2 + x + 1$ in $\mathbb{Q}[x]$.

9.43. $3x^8 - 4x^6 + 8x^5 - 10x + 6$ in $\mathbb{Q}[x]$.

9.44. $x^4 + x^2 - 6$ in $\mathbb{Q}[x]$.

9.45. $4x^3 + 3x^2 + x + 1$ in $\mathbb{Z}_5[x]$.

9.46. $x^5 + 15$ in $\mathbb{Q}[x]$.

9.47. $x^4 - 2x^3 + x^2 + 1$ in $\mathbb{R}[x]$.

9.48. Is $\mathbb{Z}[x]$ a euclidean ring when $\delta(f(x)) = \deg f(x)$ for any nonzero polynomial? Is $\mathbb{Z}[x]$ a euclidean ring with any other definition of $\delta(f(x))$?

9.49. Can you define a division algorithm in $\mathbb{R}[x, y]$? How would you divide $x^3 + 3xy + y + 4$ by $xy + y^3 + 2$?

9.50. Let L_p be the set of all linear functions $f : \mathbb{Z}_p \to \mathbb{Z}_p$ of the form $f(x) = ax + b$, where $a \neq 0$ in $\mathbb{Z}_p$. Show that $(L_p, \circ)$ is a group of order $p(p-1)$ under composition.

9.51. If p is a prime, prove that $(x - a) | (x^{p-1} - 1)$ in $\mathbb{Z}_p[x]$ for all nonzero a in $\mathbb{Z}_p$. Hence prove that

$$x^{p-1} - 1 = (x-1)(x-2)\cdots(x-p+1) \quad \text{in } \mathbb{Z}_p[x].$$

9.52. *(Wilson's theorem)* Prove that $(n-1)! \equiv -1 \mod n$ if and only if n is prime.

9.53. Prove that $\sqrt{2}/\sqrt[3]{5}$ is irrational.

9.54. Find a polynomial in $\mathbb{Q}[x]$ with $\sqrt{2} + \sqrt{3}$ as a root. Then prove that $\sqrt{2} + \sqrt{3}$ is irrational.

9.55. Is 5 irreducible in $\mathbb{Z}[i]$?

9.56. Show that $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} | a, b \in \mathbb{Z}\}$ does not have the unique factorization property.

9.57. Prove that a gaussian integer is irreducible if and only if it is an invertible element times one of the following gaussian integers:

(1) any prime p in $\mathbb{Z}$ with $p \equiv 3 \mod 4$.

(2) $1 + i$.

(3) $a + bi$, where a is positive and even, and $a^2 + b^2 = p$, for some prime p in $\mathbb{Z}$ such that $p \equiv 1 \mod 4$.

9.58. If r/s is a rational root, in its lowest terms, of a polynomial $p(x)$ with integral coefficients, show that $p(x) = (sx - r)g(x)$ for some polynomial $g(x)$ with *integral* coefficients.

9.59. Prove that r/s, in its lowest terms, cannot be a root of the integral polynomial $p(x)$ unless $(s-r)|p(1)$. This can be used to shorten the list of possible rational roots of an integral polynomial.

9.60. Let $m = m_1 m_2 \cdots m_r$ and $M_i = m/m_i$. If $\gcd(m_i, m_j) = 1$ for $i \neq j$, each of the congruences $M_i y \equiv 1 \bmod m_i$ has a solution $y \equiv b_i \bmod m_i$. Prove that the solution to the simultaneous congruences

$$x \equiv a_1 \bmod m_1, \quad x \equiv a_2 \bmod m_2, \quad \ldots, \quad x \equiv a_r \bmod m_r$$

is $x \equiv \sum_{i=1}^{r} M_i b_i a_i \bmod m$.

For Exercises 9.61 to 9.64, solve the simultaneous congruences.

9.61. $x \equiv 5 \bmod 7$
$x \equiv 4 \bmod 6$.

9.62. $x \equiv 41 \bmod 65$
$x \equiv 35 \bmod 72$.

9.63. $x \equiv 0 \bmod 2$
$x \equiv 1 \bmod 3$
$x \equiv 2 \bmod 5$.

9.64. $x \equiv 9 \bmod 12$
$x \equiv 3 \bmod 13$
$x \equiv 6 \bmod 25$.

9.65. Prove that $\det \begin{bmatrix} 320 & 461 & 5264 & 72 \\ 702 & 1008 & -967 & -44 \\ -91 & 2333 & 46 & 127 \\ 164 & -216 & 1862 & 469 \end{bmatrix}$ is nonzero.

9.66. Solve the following simultaneous equations:

$$26x - 141y = -697$$
$$55x - 112y = 202$$

(a) in $\mathbb{Z}_2$, **(b)** in $\mathbb{Z}_3$, and **(c)** in $\mathbb{Z}_5$. Then use the Chinese remainder theorem to solve them in $\mathbb{Z}$ assuming they have a pair of integral solutions between 0 and 29.

9.67. The value of $\det \begin{bmatrix} 676 & 117 & 522 \\ 375 & 65 & 290 \\ 825 & 143 & 639 \end{bmatrix}$ is positive and less than 100.

Find its value without using a calculator. (If you get tired of doing arithmetic, calculate its value mod 10 and mod 11 and then use the Chinese remainder theorem.)

9.68. The polynomial $x^3 + 5x \in \mathbb{Z}_6[x]$ has six roots. Does this contradict Theorem 9.6?

9.69. If R is an integral domain and $R[x]$ is euclidean, show that R must be a field.

9.70. Assume that R is a euclidean domain in which $\delta(a+b) \leqslant \max\{\delta(a), \delta(b)\}$ whenever a, b, and $a+b$ are all nonzero. Show that the quotient and remainder in the division algorithm are uniquely determined.

10

QUOTIENT RINGS

In this chapter we define a quotient ring in a way similar to our definition of a quotient group. The analogue of a normal subgroup is called an *ideal*, and a quotient ring consists of the set of cosets of the ring by one of its ideals. As in groups, we have a morphism theorem connecting morphisms, ideals, and quotient rings. We discover under what conditions quotient rings are fields. This will enable us to fulfill our long-range goal of extending the number systems by defining new fields using quotient rings of some familiar rings.

IDEALS AND QUOTIENT RINGS

If $(R, +, \cdot)$ is any ring and $(S, +)$ is any subgroup of the abelian group $(R, +)$, then the quotient *group* $(R/S, +)$ has already been defined. However, R/S does not have a *ring* structure induced on it by R unless S is a special kind of subgroup called an *ideal*.

A nonempty subset I of a ring R is called an **ideal** of R if the following conditions are satisfied for all $x, y \in I$ and $r \in R$:

(i) $x - y \in I$.
(ii) $x \cdot r$ and $r \cdot x \in I$.

Condition (i) implies that $(I, +)$ is a subgroup of $(R, +)$. In any ring R, R itself is an ideal, and $\{0\}$ is an ideal.

Proposition 10.1. Let a be an element of a commutative ring R. The set $\{ar | r \in R\}$ of all multiples of a is an ideal of R called the **principal ideal** generated by a. This ideal is denoted by (a).

Proof. Let $ar, as \in (a)$ and $t \in R$. Then $ar - as = a(r - s) \in (a)$ and $(ar)t = a(rt) \in (a)$. Hence (a) is an ideal of R. □

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

For example, $(n) = n\mathbb{Z}$, consisting of all integer multiples of n, is the principal ideal generated by n in $\mathbb{Z}$.

The set of all polynomials in $\mathbb{Q}[x]$ that contain $x^2 - 2$ as a factor is the principal ideal $(x^2 - 2) = \{(x^2 - 2) \cdot p(x) | p(x) \in \mathbb{Q}[x]\}$ generated by $x^2 - 2$ in $\mathbb{Q}[x]$. The set of all real polynomials that have zero constant term is the principal ideal $(x) = \{x \cdot p(x) | p(x) \in \mathbb{R}[x]\}$ generated by x in $\mathbb{R}[x]$. It is also the set of real polynomials with 0 as a root.

The set of all real polynomials, in two variables x and y, that have a zero constant term is an ideal of $\mathbb{R}[x, y]$. However, this ideal is not principal (see Exercise 10.30).

However, every ideal is principal in many commutative rings; these are called **principal ideal rings**.

Theorem 10.2. A euclidean ring is a principal ideal ring.

Proof. Let I be any ideal of the euclidean ring R. If $I = \{0\}$, then $I = (0)$, the principal ideal generated by 0. Otherwise, I contains nonzero elements. Let b be a nonzero element of I for which $\delta(b)$ is minimal. If a is any other element in I, then, by the division algorithm, there exist $q, r \in R$ such that

$$a = q \cdot b + r \quad \text{where} \quad r = 0 \text{ or } \delta(r) < \delta(b).$$

Now $r = a - q \cdot b \in I$. Since b is a nonzero element of I for which $\delta(b)$ is minimal, it follows that r must be zero and $a = q \cdot b$. Therefore, $a \in (b)$ and $I \subseteq (b)$.

Conversely, any element of (b) is of the form $q \cdot b$ for some $q \in R$, so $q \cdot b \in I$. Therefore, $I \supseteq (b)$, which proves that $I = (b)$. Hence R is a principal ideal ring. □

Corollary 10.3. $\mathbb{Z}$ is a principal ideal ring, so is $F[x]$, if F is a field.

Proof. This follows because $\mathbb{Z}$ and $F[x]$ are euclidean rings. □

Proposition 10.4. Let I be ideal of the ring R. If I contains the identity 1, then I is the entire ring R.

Proof. Let $1 \in I$ and $r \in R$. Then $r = r \cdot 1 \in I$, so $I = R$. □

Let I be any ideal in a ring R. Then $(I, +)$ is a normal subgroup of $(R, +)$, and we denote the coset of I in R that contains r by $I + r$. Hence

$$I + r = \{i + r \in R | i \in I\}.$$

The cosets of I in R are the equivalence classes under the congruence relation modulo I. We have

$$r_1 \equiv r_2 \bmod I \quad \text{if and only if} \quad r_1 - r_2 \in I.$$

By Theorem 4.18, the set of cosets $R/I = \{I + r | r \in R\}$ is an abelian group under the operation defined by

$$(I + r_1) + (I + r_2) = I + (r_1 + r_2).$$

In fact, we get a ring structure in R/I.

Theorem 10.5. Let I be an ideal in the ring R. Then the set of cosets forms a ring $(R/I, +, \cdot)$ under the operations defined by

$$(I + r_1) + (I + r_2) = I + (r_1 + r_2)$$

and

$$(I + r_1)(I + r_2) = I + (r_1 r_2).$$

This ring $(R/I, +, \cdot)$ is called the **quotient ring** (or **factor ring**) of R by I.

Proof. As mentioned above, $(R/I, +)$ is an abelian group; thus we only have to verify the axioms related to multiplication.

We first show that multiplication is well defined on cosets. Let $I + r_1' = I + r_1$ and $I + r_2' = I + r_2$, so that $r_1' - r_1 = i_1 \in I$ and $r_2' - r_2 = i_2 \in I$. Then

$$r_1' r_2' = (i_1 + r_1)(i_2 + r_2) = i_1 i_2 + r_1 i_2 + i_1 r_2 + r_1 r_2.$$

Now, since I is an ideal, $i_1 i_2$, $r_1 i_2$ and $i_1 r_2 \in I$. Hence $r_1' r_2' - r_1 r_2 \in I$, so $I + r_1' r_2' = I + r_1 r_2$, which shows that multiplication is well defined on R/I.

Multiplication is associative and distributive over addition. If $r_1, r_2, r_3 \in R$, then

$$\begin{aligned}(I + r_1)\{(I + r_2)(I + r_3)\} &= (I + r_1)(I + r_2 r_3) = I + r_1(r_2 r_3) = I + (r_1 r_2) r_3 \\ &= (I + r_1 r_2)(I + r_3) = \{(I + r_1)(I + r_2)\}(I + r_3).\end{aligned}$$

Also,

$$\begin{aligned}(I + r_1)\{(I + r_2) + (I + r_3)\} &= (I + r_1)\{I + (r_2 + r_3)\} = I + r_1(r_2 + r_3) \\ &= I + (r_1 r_2 + r_1 r_3) = (I + r_1 r_2) + (I + r_1 r_3) \\ &= \{(I + r_1)(I + r_2)\} + \{(I + r_1)(I + r_3)\}.\end{aligned}$$

The other distributive law can be proved similarly. The multiplicative identity is $I + 1$. Hence $(R/I, +, \cdot)$ is a ring. □

TABLE 10.1. Quotient Ring $\mathbb{Z}_6/\{0,2,4\}$

$+$	I	$I+1$
I	I	$I+1$
$I+1$	$I+1$	I

$\cdot$	I	$I+1$
I	I	I
$I+1$	I	$I+1$

For example, the quotient ring of $\mathbb{Z}$ by (n) is $\mathbb{Z}/(n) = \mathbb{Z}_n$, the ring of integers modulo n. A coset $(n) + r = \{nz + r | z \in \mathbb{Z}\}$ is the equivalent class modulo n containing r.

If R is commutative, so is the quotient ring R/I, because

$$(I + r_1)(I + r_2) = I + r_1 r_2 = I + r_2 r_1 = (I + r_2)(I + r_1).$$

Example 10.6. If $I = \{0, 2, 4\}$ is the ideal generated by 2 in $\mathbb{Z}_6$, find the tables for the quotient ring $\mathbb{Z}_6/I$.

Solution. There are two cosets of $\mathbb{Z}_6$ by I: namely, $I = \{0, 2, 4\}$ and $I + 1 = \{1, 3, 5\}$. Hence

$$\mathbb{Z}_6/I = \{I, I + 1\}.$$

The addition and multiplication tables given in Table 10.1 show that the quotient ring $\mathbb{Z}_6/I$ is isomorphic to $\mathbb{Z}_2$. □

COMPUTATIONS IN QUOTIENT RINGS

If F is a field, the quotient rings of the polynomial ring $F[x]$ form an important class of rings that will be used to construct new fields. Recall that $F[x]$ is a principal ideal ring, so that any quotient ring is of the form $F[x]/(p(x))$, for some polynomial $p(x) \in F[x]$. We now look at the structure of such a quotient ring.

The elements of the ring $F[x]/(p(x))$ are equivalence classes under the relation on $F[x]$ defined by

$$f(x) \equiv g(x) \bmod(p(x)) \quad \text{if and only if} \quad f(x) - g(x) \in (p(x)).$$

Lemma 10.7. $f(x) \equiv g(x) \bmod(p(x))$ if and only if $f(x)$ and $g(x)$ have the same remainder when divided by $p(x)$.

Proof. Let $f(x) = q(x) \cdot p(x) + r(x)$ and $g(x) = s(x) \cdot p(x) + t(x)$, where $r(x)$ and $t(x)$ are zero or have degrees less than that of $p(x)$. Now the lemma follows because the following statements are equivalent:

(i) $f(x) \equiv g(x) \bmod(p(x))$.
(ii) $f(x) - g(x) \in (p(x))$.
(iii) $p(x) | f(x) - g(x)$.

(iv) $p(x)|[\{q(x) - s(x)\} \cdot p(x) + (r(x) - t(x))]$.
(v) $p(x)|[r(x) - t(x)]$.
(vi) $r(x) = t(x)$. □

Hence every coset of $F[x]$ by $(p(x))$ contains the zero polynomial or a polynomial of degree less than that of $p(x)$.

Theorem 10.8. If F is a field, let P be the ideal $(p(x))$ in $F[x]$ generated by the polynomial $p(x)$ of degree $n > 0$. The different elements of $F[x]/(p(x))$ are precisely those of the form

$$P + a_0 + a_1x + \cdots + a_{n-1}x^{n-1} \quad \text{where} \quad a_0, a_1, \ldots, a_{n-1} \in F.$$

Proof. Let $P + f(x)$ be any element of $F[x]/(p(x))$ and let $r(x)$ be the remainder when $f(x)$ is divided by $p(x)$. Then, by Lemma 10.7, $P + f(x) = P + r(x)$, which is of the required form.

Suppose that $P + r(x) = P + t(x)$, where $r(x)$ and $t(x)$ are zero or have degree less than n. Then

$$r(x) \equiv t(x) \bmod(p(x)),$$

and by Lemma 10.7, $r(x) = t(x)$. □

Example 10.9. Write down the tables for $\mathbb{Z}_2[x]/(x^2 + x + 1)$.

Solution. Let $P = (x^2 + x + 1)$, so that

$$\begin{aligned}\mathbb{Z}_2[x]/(x^2 + x + 1) &= \{P + a_0 + a_1x | a_0, a_1 \in \mathbb{Z}_2\} \\ &= \{P, P + 1, P + x, P + x + 1\}.\end{aligned}$$

The tables for the quotient ring are given in Table 10.2. The addition table is straightforward to calculate. Multiplication is computed as follows:

$$(P + x)^2 = P + x^2 = P + (x^2 + x + 1) + (x + 1) = P + x + 1$$

and

$$(P + x)(P + x + 1) = P + x^2 + x = P + (x^2 + x + 1) + 1 = P + 1. \quad \square$$

Example 10.10. Let $P = x^2 - 2$ be the principal ideal of $\mathbb{Q}[x]$ generated by $x^2 - 2$. Find the sum and product of $P + 3x + 4$ and $P + 5x - 6$ in the ring $\mathbb{Q}[x]/(x^2 - 2) = \{P + a_0 + a_1x | a_0, a_1 \in \mathbb{Q}\}$.

Solution. $(P + 3x + 4) + (P + 5x - 6) = P + (3x + 4) + (5x - 6) = P + 8x - 2$. $(P + 3x + 4)(P + 5x - 6) = P + (3x + 4)(5x - 6) = P + 15x^2 + 2x - 24$. By the division algorithm, $15x^2 + 2x - 24 = 15(x^2 - 2) + 2x + 6$. Hence, by Lemma 10.7, $P + 15x^2 + 2x - 24 = P + 2x + 6$. □

TABLE 10.2. Ring $\mathbb{Z}_2[x]/(x^2+x+1)$

$+$	P	$P+1$	$P+x$	$P+x+1$
P	P	$P+1$	$P+x$	$P+x+1$
$P+1$	$P+1$	P	$P+x+1$	$P+x$
$P+x$	$P+x$	$P+x+1$	P	$P+1$
$P+x+1$	$P+x+1$	$P+x$	$P+1$	P

$\cdot$	P	$P+1$	$P+x$	$P+x+1$
P	P	P	P	P
$P+1$	P	$P+1$	$P+x$	$P+x+1$
$P+x$	P	$P+x$	$P+x+1$	$P+1$
$P+x+1$	P	$P+x+1$	$P+1$	$P+x$

There are often easier ways of finding the remainder of $f(x)$ when divided by $p(x)$ than by applying the division algorithm directly. If $\deg p(x) = n$ and $P = (p(x))$, the problem of finding the remainder reduces to the problem of finding a polynomial $r(x)$ of degree less than n such that $f(x) \equiv r(x) \bmod P$. This can often be solved by manipulating congruences, using the fact that $p(x) \equiv 0 \bmod P$.

Consider Example 10.10, in which P is the ideal generated by $x^2 - 2$. Then $x^2 - 2 \equiv 0 \bmod P$ and $x^2 \equiv 2 \bmod P$. Hence, in any congruence modulo P, we can always replace x^2 by 2. For example,

$$\begin{aligned} 15x^2 + 2x - 24 &\equiv 15(2) + 2x - 24 \bmod P \\ &\equiv 2x + 6 \bmod P, \end{aligned}$$

so $P + 15x^2 + 2x - 24 = P + 2x + 6$.

In Example 10.9, $P = (x^2 + x + 1)$, so $x^2 + x + 1 \equiv 0 \bmod P$ and $x^2 \equiv x + 1 \bmod P$. (Remember $+1 = -1$ in $\mathbb{Z}_2$.) Therefore, in multiplying two elements in $\mathbb{Z}_2[x]/P$, we can always replace x^2 by $x + 1$. For example,

$$P + x^2 = P + x + 1 \quad \text{and} \quad P + x(x+1) = P + x^2 + x = P + 1.$$

We have usually written the elements of $\mathbb{Z}_n = \mathbb{Z}/(n)$ simply as $0, 1, \ldots, n-1$ instead of as $[0], [1], \ldots, [n-1]$ or as $(n)+0, (n)+1, \ldots, (n)+n-1$. In a similar way, when there is no confusion, we henceforth write the elements of $F[x]/(p(x))$ simply as $a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$ instead of $(p(x)) + a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$.

MORPHISM THEOREM

Proposition 10.11. If $f: R \to S$ is a ring morphism, then $\operatorname{Ker} f$ is an ideal of R.

Proof. Since any ring morphism is a group morphism, it follows from Proposition 4.23 that $\operatorname{Ker} f$ is a subgroup of $(R, +)$. If $x \in \operatorname{Ker} f$ and $r \in R$, then

$f(xr) = f(x)f(r) = 0 \cdot f(r) = 0$ and $xr \in \operatorname{Ker} f$. Similarly, $rx \in \operatorname{Ker} f$, so $\operatorname{Ker} f$ is an ideal of R. □

Furthermore, any ideal I of a ring R is the kernel of a morphism, for example, the ring morphism $\pi: R \to R/I$ defined by $\pi(r) = I + r$.

The image of a morphism $f: R \to S$ can easily be verified to be a subring of S.

Theorem 10.12. Morphism Theorem for Rings. If $f: R \to S$ is a ring morphism, then $R/\operatorname{Ker} f$ is isomorphic to $\operatorname{Im} f$.

This result is also known as the **first isomorphism theorem for rings**; the second and third isomorphism theorems are given in Exercises 10.19 and 10.20.

Proof. Let $K = \operatorname{Ker} f$. It follows from the morphism theorem for groups (Theorem 4.23), that $\psi: R/K \to \operatorname{Im} f$, defined by $\psi(K + r) = f(r)$, is a group isomorphism. Hence we need only prove that ψ is a ring morphism. We have

$$\begin{aligned}\psi\{(K + r)(K + s)\} &= \psi\{K + rs\} = f(rs) = f(r)f(s)\\ &= \psi(K + r)\psi(K + s).\end{aligned}$$ □

Example 10.13. Prove that $\mathbb{Q}[x]/(x^2 - 2) \cong \mathbb{Q}(\sqrt{2})$.

Solution. Consider the ring morphism $\psi: \mathbb{Q}[x] \to \mathbb{R}$ defined by $\psi(f(x)) = f(\sqrt{2})$ in Example 9.24. The kernel is the set of polynomials containing $x^2 - 2$ as a factor, that is, the principal ideal $(x^2 - 2)$. The image of ψ is $\mathbb{Q}(\sqrt{2})$ so by the morphism theorem for rings, $\mathbb{Q}[x]/(x^2 - 2) \cong \mathbb{Q}(\sqrt{2})$. □

In this isomorphism, the element $a_0 + a_1x \in \mathbb{Q}[x]/(x^2 - 2)$ is mapped to $a_0 + a_1\sqrt{2} \in \mathbb{Q}(\sqrt{2})$. Addition and multiplication of the elements $a_0 + a_1x$ and $b_0 + b_1x$ in $\mathbb{Q}[x]/(x^2 - 2)$ correspond to the addition and multiplication of the real numbers $a_0 + a_1\sqrt{2}$ and $b_0 + b_1\sqrt{2}$.

Example 10.14. Prove that $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$.

Solution. Define the ring morphism $\psi: \mathbb{R}[x] \to \mathbb{C}$ by $\psi(f(x)) = f(i)$, where $i = \sqrt{-1}$. Any polynomial in $\operatorname{Ker}\psi$ has i as a root, and therefore, by Theorem 9.22, also has $-i$ as a root and contains the factor $x^2 + 1$. Hence $\operatorname{Ker}\psi = (x^2 + 1)$.

Now $\psi(a + bx) = a + ib$; thus ψ is surjective. By the morphism theorem for rings, $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$. □

QUOTIENT POLYNOMIAL RINGS THAT ARE FIELDS

We now determine when a quotient of a polynomial ring is a field. This result allows us to construct many new fields.

Theorem 10.15. Let a be an element of the euclidean ring R. The quotient ring $R/(a)$ is a field if and only if a is irreducible in R.

Proof. Suppose that a is an irreducible element of R and let $(a) + b$ be a nonzero element of $R/(a)$. Then b is not a multiple of a, and since a is irreducible, $\gcd(a, b) = 1$. By Theorem 9.9, there exist $s, t \in R$ such that

$$sa + tb = 1.$$

Now $sa \in (a)$, so $[(a) + t] \cdot [(a) + b] = (a) + 1$, the identity of $R/(a)$. Hence $(a) + t$ is the inverse of $(a) + b$ in $R/(a)$ and $R/(a)$ is a field.

Now suppose that a is not irreducible in R so that there exist elements s and t, which are not invertible, with $st = a$. By Lemma 9.17, $\delta(s) < \delta(st) = \delta(a)$ and $\delta(t) < \delta(st) = \delta(a)$. Hence s is not divisible by a, and $s \notin (a)$. Similarly, $t \notin (a)$, and neither $(a) + s$ nor $(a) + t$ is the zero element of $R/(a)$. However,

$$[(a) + s] \cdot [(a) + t] = (a) + st = (a), \quad \text{the zero element of } R/(a).$$

Therefore, the ring $R/(a)$ has zero divisors and cannot possibly be a field. □

For example, in the quotient ring $\mathbb{Q}[x]/P$, where $P = (x^2 - 1)$, the elements $P + x + 1$ and $P + x - 1$ are zero divisors because

$$(P + x + 1) \cdot (P + x - 1) = P + x^2 - 1 = P, \quad \text{the zero element.}$$

Corollary 10.16. $\mathbb{Z}_p = \mathbb{Z}/(p)$ is a field if and only if p is prime.

Proof. This result, which we proved in Theorem 8.11, follows from Theorem 10.15 because the irreducible elements in $\mathbb{Z}$ are the primes (and their negatives). □

Another particular case of Theorem 10.15 is the following important theorem.

Theorem 10.17. The ring $F[x]/(p(x))$ is a field if and only if $p(x)$ is irreducible over the field F. Furthermore, the ring $F[x]/(p(x))$ always contains a subring isomorphic to the field F.

Proof. The first part of the theorem is just Theorem 10.15. Let $F = \{(p(x)) + r | r \in F\}$. This can be verified to be a subring of $F[x]/(p(x))$, which is isomorphic to the field F by the isomorphism that takes $r \in F$ to $(p(x)) + r \in F[x]/(p(x))$. □

Example 10.18. Show that $\mathbb{Z}_2[x]/(x^2 + x + 1)$ is a field with four elements.

Solution. We showed in Example 9.34 that $x^2 + x + 1$ is irreducible over $\mathbb{Z}_2$ and in Example 10.9 that the quotient ring has four elements. Hence the quotient ring is a field containing four elements. Its tables are given in Table 10.2. □

Example 10.19. Write down the multiplication table for the field $\mathbb{Z}_3[x]/(x^2+1)$.

Solution. If $x = 0$, 1, or 2 in $\mathbb{Z}_3$, then $x^2 + 1 = 1$, 2, or 2; thus, by the factor theorem, $x^2 + 1$ has no linear factors. Hence $x^2 + 1$ is irreducible over $\mathbb{Z}_3$ and, by Theorem 10.17, the quotient ring $\mathbb{Z}_3[x]/(x^2 + 1)$ is a field. By Theorem 10.8, the elements of this field can be written as

$$\mathbb{Z}_3[x]/(x^2+1) = \{a_0 + a_1x | a_0, a_1 \in \mathbb{Z}_3\}.$$

Hence the field contains nine elements. Its multiplication table is given in Table 10.3. This can be calculated by multiplying the polynomials in $\mathbb{Z}_3[x]$ and replacing x^2 by -1 or 2, since $x^2 \equiv -1 \equiv 2 \text{mod } (x^2 + 1)$. □

Example 10.20. Show that $\mathbb{Q}[x]/(x^3 - 5) = \{a_0 + a_1x + a_2x^2 | a_i \in \mathbb{Q}\}$ is a field and find the inverse of the element $x + 1$.

Solution. By the rational roots theorem (Theorem 9.25), $(x^3 - 5)$ has no linear factors and hence is irreducible over $\mathbb{Q}$. Therefore, by Theorem 10.17, $\mathbb{Q}[x]/(x^3 - 5)$ is a field.

If $s(x)$ is the inverse of $x + 1$, then $(x + 1)s(x) \equiv 1 \text{ mod}(x^3 - 5)$; that is, $(x + 1)s(x) + (x^3 - 5)t(x) = 1$ for some $t(x) \in \mathbb{Q}[x]$.

We can find such polynomials $s(x)$ and $t(x)$ by the euclidean algorithm. We have (see below)

$$x^3 - 5 = (x^2 - x + 1)(x + 1) - 6,$$

so

$$6 \equiv (x^2 - x + 1)(x + 1) \text{ mod}(x^3 - 5)$$

and

$$1 \equiv \tfrac{1}{6}(x^2 - x + 1)(x + 1) \text{ mod}(x^3 - 5).$$

TABLE 10.3. Multiplication in $\mathbb{Z}_3[x]/(x^2 + 1)$

·	0	1	2	x	$x+1$	$x+2$	2x	$2x+1$	$2x+2$
0	0	0	0	0	0	0	0	0	0
1	0	1	2	x	$x+1$	$x+2$	$2x$	$2x+1$	$2x+2$
2	0	2	1	$2x$	$2x+2$	$2x+1$	x	$x+2$	$x+1$
x	0	x	$2x$	2	$x+2$	$2x+2$	1	$x+1$	$2x+1$
$x+1$	0	$x+1$	$2x+2$	$x+2$	$2x$	1	$2x+1$	2	x
$x+2$	0	$x+2$	$2x+1$	$2x+2$	1	x	$x+1$	$2x$	2
$2x$	0	$2x$	x	1	$2x+1$	$x+1$	2	$2x+2$	$x+2$
$2x+1$	0	$2x+1$	$x+2$	$x+1$	2	$2x$	$2x+2$	x	1
$2x+2$	0	$2x+2$	$x+1$	$2x+1$	x	2	$x+2$	1	$2x$

Hence $(x+1)^{-1} = \frac{1}{6}x^2 - \frac{1}{6}x + \frac{1}{6}$ in $\mathbb{Q}[x]/(x^3-5)$.

$$\begin{array}{r|l} & x^2 - x + 1 \\ \hline x+1 & x^3 + 0 + 0 - 5 \\ & x^3 + x^2 \\ \hline & -x^2 - 5 \\ & -x^2 - x \\ \hline & \quad x - 5 \\ & \quad x + 1 \\ \hline & \quad -6 \end{array}$$

□

Example 10.21. Show that $\mathbb{Z}_3[x]/(x^3+2x+1)$ is a field with 27 elements and find the inverse of the element x^2.

Solution. If $x = 0$, 1, or 2 in $\mathbb{Z}_3$, then $x^3 + 2x + 1 = 1$; hence $x^3 + 2x + 1$ has no linear factors and is irreducible. Therefore,

$$\mathbb{Z}_3[x]/(x^3+2x+1) = \{a_0 + a_1x + a_2x^2 | a_i \in \mathbb{Z}_3\}$$

is a field that has $3^3 = 27$ elements.

As in Example 10.20, to find the inverse of x^2, we apply the euclidean algorithm to $x^3 + 2x + 1$ and x^2 in $\mathbb{Z}_3[x]$.

We have $x^3 + 2x + 1 = x(x^2) + (2x+1)$ and $x^2 = (2x+2)(2x+1) + 1$. Hence

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

so

$$1 \equiv x^2(2x^2+2x+1) \bmod (x^3+2x+1)$$

and the inverse of x^2 in $\mathbb{Z}_3[x]/(x^3+2x+1)$ is $2x^2+2x+1$.

$$\begin{array}{r|l} & x \\ \hline x^2 & x^3 + 0 + 2x + 1 \\ & x^3 \\ \hline & \quad 2x + 1 \end{array} \qquad \begin{array}{r|l} & 2x + 2 \\ \hline 2x+1 & x^2 + 0 + 0 \\ & x^2 + 2x \\ \hline & \quad x + 0 \\ & \quad x + 2 \\ \hline & \quad 1 \end{array}$$

□

We cannot use Theorem 10.15 directly on a field to obtain any new quotient fields, because the only ideals of a field are the zero ideal and the entire field. In fact, the following result shows that a field can be characterized by its ideals.

Theorem 10.22. The nontrivial commutative ring R is a field if and only if (0) and R are its only ideals.

Proof. Let I be an ideal in the field R. Suppose that $I \neq (0)$, so that there is a nonzero element $a \in I$. Since $a^{-1} \in R$, $a \cdot a^{-1} = 1 \in I$. Therefore, by Proposition 10.4, $I = R$. Hence R has only trivial ideals.

Conversely, suppose that (0) and R are the only ideals in the ring R. Let a be a nonzero element of R and consider (a) the principal ideal generated by a. Since $1 \cdot a \in (a)$, $(a) \neq (0)$, and hence $(a) = R$. Hence $1 \in R = (a)$, so there must exist some $b \in R$ such that $a \cdot b = 1$. Therefore, $b = a^{-1}$ and R is a field. □

EXERCISES

For Exercises 10.1 to 10.6, find all the ideals in the rings.

10.1. $\mathbb{Z}_2 \times \mathbb{Z}_2$. **10.2.** $\mathbb{Z}_{18}$. **10.3.** $\mathbb{Q}$.

10.4. $\mathbb{Z}_7$. **10.5.** $\mathbb{C}[x]$. **10.6.** $\mathbb{Z}[i]$.

For Exercises 10.7 to 10.10, construct addition and multiplication tables for the rings. Find all the zero divisors in each ring. Which of these rings are fields?

10.7. $\mathbb{Z}_6/(3)$. **10.8.** $\mathbb{Z}_2[x]/(x^3 + 1)$.

10.9. $\mathbb{Z}_3 \times \mathbb{Z}_3/((1, 2))$. **10.10.** $\mathbb{Z}_3[x]/(x^2 + 2x + 2)$.

For Exercises 10.11 to 10.14, compute the sum and product of the elements in the given quotient rings.

10.11. $3x + 4$ and $5x - 2$ in $\mathbb{Q}[x]/(x^2 - 7)$.

10.12. $x^2 + 3x + 1$ and $-2x^2 + 4$ in $\mathbb{Q}[x]/(x^3 + 2)$.

10.13. $x^2 + 1$ and $x + 1$ in $\mathbb{Z}_2[x]/(x^3 + x + 1)$.

10.14. $ax + b$ and $cx + d$ in $\mathbb{R}[x]/(x^2 + 1)$, where $a, b, c, d \in \mathbb{R}$.

10.15. If U and V are ideals in a ring R, prove that $U \cap V$ is also an ideal in R.

10.16. Show, by example, that if U and V are ideals in a ring R, then $U \cup V$ is not necessarily an ideal in R. But prove that $U + V = \{u + v | u \in U, v \in V\}$ is always an ideal in R.

10.17. Find a generator of the following ideals in the given ring and prove a general result for the intersection of two ideals in a principal ideal ring.

(a) $(2) \cap (3)$ in $\mathbb{Z}$. **(b)** $(12) \cap (18)$ in $\mathbb{Z}$.

(c) $(x^2 - 1) \cap (x + 1)$ in $\mathbb{Q}[x]$.

10.18. Find a generator of the following ideals in the given ring and prove a general result for the sum of two ideals in a principal ideal ring.
(a) $(2) + (3)$ in $\mathbb{Z}$. **(b)** $(9) + (12)$ in $\mathbb{Z}$.
(c) $(x^2 + x + 1) + (x^2 + 1)$ in $\mathbb{Z}_2[x]$.

10.19. (**Second isomorphism theorem for rings**) If I and J are ideals of the ring R, prove that

$$I/(I \cap J) \cong (I + J)/J.$$

10.20. (**Third isomorphism theorem for rings**) Let I and J be two ideals of the ring R, with $J \subseteq I$. Prove that I/J is an ideal of R/J and that

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

For Exercises 10.21 to 10.29, prove the isomorphisms.

10.21. $\mathbb{R}[x]/(x^2 + 5) \cong \mathbb{C}$.

10.22. $\mathbb{Z}[x]/(x^2 + 1) \cong \mathbb{Z}[i] = \{a + ib | a, b \in \mathbb{Z}\}$.

10.23. $\mathbb{Q}[x]/(x^2 - 7) \cong \mathbb{Q}(\sqrt{7}) = \{a + b\sqrt{7} | a, b \in \mathbb{Q}\}$.

10.24. $\mathbb{Z}[x]/(2x - 1) \cong \{a/b \in \mathbb{Q} | a \in \mathbb{Z}, b = 2^r, r \geqslant 0\}$, a subring of $\mathbb{Q}$.

10.25. $\mathbb{Z}_{14}/(7) \cong \mathbb{Z}_7$. **10.26.** $\mathbb{Z}_{14}/(2) \cong \mathbb{Z}_2$.

10.27. $\mathbb{R}[x, y]/(x + y) \cong \mathbb{R}[y]$. **10.28.** $(R \times S)/((1, 0)) \cong S$.

10.29. $\mathcal{P}(X)/\mathcal{P}(X - Y) \cong \mathcal{P}(Y)$, where Y is a subset of X and the operations in these boolean rings are symmetric difference and intersection.

10.30. Let I be the set of all polynomials with no constant term in $\mathbb{R}[x, y]$. Find a ring morphism from $\mathbb{R}[x, y]$ to $\mathbb{R}$ whose kernel is the ideal I. Prove that I is not a principal ideal.

10.31. Let $I = \{p(x) \in \mathbb{Z}[x] | \ 5 | p(0)\}$. Prove that I is an ideal of $\mathbb{Z}[x]$ by finding a ring morphism from $\mathbb{Z}[x]$ to $\mathbb{Z}_5$ with kernel I. Prove that I is not a principal ideal.

10.32. Let $I \subseteq \mathcal{P}(X)$ with the property that, if $A \in I$, then all the subsets of A are in I, and also if A and B are disjoint sets in I, then $A \cup B \in I$. Prove that I is an ideal in the boolean ring $(\mathcal{P}(X), \Delta, \cap)$.

10.33. Is $\{p(x) \in \mathbb{Q}[x] | p(0) = 3\}$ an ideal of $\mathbb{Q}[x]$?

10.34. Is $\left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix} \in M_2(\mathbb{Z}) | a, b \in \mathbb{Z} \right\}$ an ideal of $M_2(\mathbb{Z})$?

10.35. What is the smallest ideal in $M_2(\mathbb{Z})$ containing $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$?

10.36. Let a, b be elements of a euclidean ring R. Prove that

$$(a) \subseteq (b) \quad \text{if and only if} \quad b | a.$$

For the rings in Exercises 10.37 and 10.38, find all the ideals and draw the poset diagrams of the ideals under inclusion.

10.37. $\mathbb{Z}_8$. **10.38.** $\mathbb{Z}_{20}$.

Which of the elements in Exercises 10.39 to 10.46 are irreducible in the given ring? If an element is irreducible, find the corresponding quotient field modulo the ideal generated by that element.

10.39. 11 in $\mathbb{Z}$.

10.40. 10 in $\mathbb{Z}$.

10.41. $x^2 - 2$ in $\mathbb{R}[x]$.

10.42. $x^3 + x^2 + 2$ in $\mathbb{Z}_3[x]$.

10.43. $x^4 - 2$ in $\mathbb{Q}[x]$.

10.44. $x^7 + 4x^3 - 3ix + 1$ in $\mathbb{C}[x]$.

10.45. $x^2 - 3$ in $\mathbb{Q}(\sqrt{2})[x]$.

10.46. $3x^5 - 4x^3 + 2$ in $\mathbb{Q}[x]$.

Which of the rings in Exercises 10.47 to 10.56 are fields? Give reasons.

10.47. $\mathbb{Z}_2 \times \mathbb{Z}_2$.

10.48. $\mathbb{Z}_4$.

10.49. $\mathbb{Z}_{17}$.

10.50. $\mathbb{R}^3$.

10.51. $\mathbb{Q}[x]/(x^3 - 3)$.

10.52. $\mathbb{Z}_7[x]/(x^2 + 1)$.

10.53. $\mathbb{Z}_5[x]/(x^2 + 1)$.

10.54. $\mathbb{R}[x]/(x^2 + 7)$.

10.55. $\mathbb{Q}(\sqrt[4]{11}) = \{a + b11^{\frac{1}{4}} + c11^{\frac{1}{2}} + d11^{\frac{3}{4}} | a, b, c, d \in \mathbb{Q}\}$.

10.56. $M_n(\mathbb{R})$.

10.57. An ideal $I \neq R$ is said to be a **maximal ideal** in the commutative ring R if, whenever U is an ideal of R such that $I \subseteq U \subseteq R$, then $U = I$ or $U = R$. Show that the nonzero ideal (a) of a euclidean ring R is maximal if and only if a is irreducible in R.

10.58. If I is an ideal in a commutative ring R, prove that R/I is a field if and only if I is a maximal ideal of R.

10.59. Find all the ideals in the ring of formal power series, $\mathbb{R}[[x]]$. Which of the ideals are maximal?

10.60. Let $C[0, 1] = \{f : [0, 1] \to \mathbb{R} | f \text{ is continuous}\}$, the ring of real-valued continuous functions on the interval [0, 1]. Prove that $I_a = \{f \in C[0, 1] | f(a) = 0\}$ is a maximal ideal in $C[0, 1]$ for each $a \in [0, 1]$. (Every maximal ideal is, in fact, of this form, but this is much harder to prove.)

10.61. If P is an ideal in a commutative ring R, show that R/P is an integral domain if and only if $ab \in P$ can only happen if $a \in P$ or $b \in P$. The ideal P is called a **prime ideal** in this case.

10.62. Let R be a commutative ring. An element $a \in R$ is called **nilpotent** if $a^n = 0$ for some $n \geqslant 0$ in $\mathbb{Z}$. The set $N(R)$ of all nilpotents in R is called the **radical** of R.

(a) Show that $N(R)$ is an ideal of R (the binomial theorem is useful).

(b) Show that $N[R/N(R)] = \{N(R)\}$.

(c) Show that $N(R)$ is contained in the intersection of all prime ideals of R (see Exercise 10.61). In fact, $N(R)$ *equals* the intersection of all prime ideals of R.

10.63. A commutative ring R is called **local** if the set $J(R)$ of all non-invertible elements forms an ideal of R.

(a) Show that every field is local as is $\mathbb{Z}_{p^n}$ for each prime p and $n \geqslant 1$, but that $\mathbb{Z}_6$ is *not* local.

(b) Show that $\mathbb{Z}_{(p)} = \left\{ \frac{r}{s} \in \mathbb{Q} | p \text{ does not divide } s \right\}$ is a local subring of $\mathbb{Q}$ for each prime p.

(c) If R is local show that $R/J(R)$ is a field.

(d) If $R/N(R)$ is a field, show that R is local (if a is nilpotent, then $1 - a$ is invertible).

10.64. If R is a (possibly noncommutative) ring, an additive subgroup L of R is called a **left ideal** if $rx \in L$ for all $r \in R$ and $x \in L$. Show that the only left ideals of R are $\{0\}$ and R if and only if every nonzero element of R has an inverse (then R is called a **skew field**).

10.65. If F is a field, show that $R = M_2(F)$ has exactly two ideals: 0 and R. (Because of this, R is called a **simple ring**.) Conclude that Theorem 10.22 fails if the ring is not commutative.

11

FIELD EXTENSIONS

We proved in Chapter 10 that if $p(x)$ is an irreducible polynomial over the field F, the quotient ring $K = F[x]/(p(x))$ is a field. This field K contains a subring isomorphic to F; thus K can be considered to be an extension of the field F. We show that the polynomial $p(x)$ now has a root α in this extension field K, even though $p(x)$ was irreducible over F. We say that K can be obtained from F by adjoining the root α. We can construct the complex numbers $\mathbb{C}$ in this way, by adjoining a root of $x^2 + 1$ to the real numbers $\mathbb{R}$.

Another important achievement is the construction of a finite field with p^n elements for each prime p. Such a field is called a *Galois field of order* p^n and is denoted by $GF(p^n)$. We show how this field can be constructed as a quotient ring of the polynomial ring $\mathbb{Z}_p[x]$, by an irreducible polynomial of degree n.

FIELD EXTENSIONS

A **subfield** of a field K is a subring F that is also a field. In this case, the field K is called an **extension of the field** F. For example, $\mathbb{Q}$ is a subfield of $\mathbb{R}$; thus $\mathbb{R}$ is an extension of the field $\mathbb{Q}$.

Example 11.1. Let $p(x)$ be a polynomial of degree n irreducible over the field F, so that the quotient ring

$$K = F[x]/(p(x)) = \{a_0 + a_1x + \cdots + a_{n-1}x^{n-1} | a_i \in F\}$$

is a field. Then K is an extension field of F.

Solution. This follows from Theorem 10.17 when we identify the coset $(p(x)) + a_0$ containing the constant term a_0 with the element a_0 of F. □

Proposition 11.2. Let K be an extension field of F. Then K is a vector space over F.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

Proof. K is an abelian group under addition. Elements of K can be multiplied by elements of F. This multiplication satisfies the following properties:

(i) If 1 is the identity element of F then $1k = k$ for all $k \in K$.
(ii) If $\lambda \in F$ and $k, l \in K$, then $\lambda(k + l) = \lambda k + \lambda l$.
(iii) If $\lambda, \mu \in F$ and $k \in K$, then $(\lambda + \mu)k = \lambda k + \mu K$.
(iv) If $\lambda, \mu \in F$ and $k \in K$, then $(\lambda\mu)k = \lambda(\mu k)$.

Hence K is a vector space over F. □

The fact that a field extension K is a vector space over F tells us much about the structure of K. The elements of K can be written uniquely as a linear combination of certain elements called *basis elements*. Furthermore, if the vector space K has finite dimension n over the field F, there will be n basis elements, and the construction of K is particularly simple.

The **degree** of the extension K of the field F, written $[K : F]$, is the dimension of K as a vector space over F. The field K is called a **finite extension** if $[K : F]$ is finite.

Example 11.3. $[\mathbb{C} : \mathbb{R}] = 2$.

Solution. $\mathbb{C} = \{a + ib | a, b \in \mathbb{R}\}$; therefore, 1 and i span the vector space $\mathbb{C}$ over $\mathbb{R}$. Now 1 and i are linearly independent since, if $\lambda, \mu \in \mathbb{R}$, then $\lambda 1 + \mu i = 0$ implies that $\lambda = \mu = 0$. Hence $\{1, i\}$ is a basis for $\mathbb{C}$ over $\mathbb{R}$ and $[\mathbb{C} : \mathbb{R}] = 2$. □

Example 11.4. If $K = \mathbb{Z}_5[x]/(x^3 + x + 1)$, then $[K : \mathbb{Z}_5] = 3$.

Solution. $\{1, x, x^2\}$ is a basis for K over $\mathbb{Z}_5$ because by Theorem 10.8, every element of K can be written uniquely as the coset containing $a_0 + a_1 x + a_2 x^2$, where $a_i \in \mathbb{Z}_5$. Hence $[K : \mathbb{Z}_5] = 3$. □

Example 11.4 is a special case of the following theorem.

Theorem 11.5. If $p(x)$ is an irreducible polynomial of degree n over the field F, and $K = F[x]/(p(x))$, then $[K : F] = n$.

Proof. By Theorem 10.8, $K = \{a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} | a_i \in F\}$, and such expressions for the elements of K are unique. Hence $\{1, x, x^2, \ldots, x^{n-1}\}$ is a basis for K over F, and $[K : F] = n$. □

Theorem 11.6. Let L be a finite extension of K and K a finite extension of F. Then L is a finite extension of F and $[L : F] = [L : K][K : F]$.

Proof. We have three fields, F, K, L, with $L \supseteq K \supseteq F$. We prove the theorem by taking bases for L over K, and K over F, and constructing a basis for L over F.

Let $[L : K] = m$ and let $\{u_1, \ldots, u_m\}$ be a basis for L over K. Let $[K : F] = n$ and let $\{v_1, \ldots, v_n\}$ be a basis for K over F. We show that

$$\mathcal{B} = \{v_j u_i | i = 1, \ldots, m, \ \ j = 1, \ldots, n\} \text{ is a basis for } L \text{ over } F.$$

If $x \in L$, then $x = \sum_{i=1}^{m} \lambda_i u_i$, for some $\lambda_i \in K$. Now each element λ_i can be written as $\lambda_i = \sum_{j=1}^{n} \mu_{ij} v_j$, for some $\mu_{ij} \in F$. Hence $x = \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} v_j u_i$, and $\mathcal{B}$ spans L over F.

Now suppose that $\sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij} v_j u_i = 0$, where $\mu_{ij} \in F$. Then, since $u_1, \ldots, u_m$ are linearly independent over K, it follows that $\sum_{j=1}^{n} \mu_{ij} v_j = 0$ for each $i = 1, \ldots, m$. But $v_1, \ldots, v_n$ are linearly independent over F so $\mu_{ij} = 0$ for each i and each j.

Hence the elements of $\mathcal{B}$ are linearly independent, and $\mathcal{B}$ is a basis for L over F. Therefore, $[L : F] = m \cdot n = [L : K][K : F]$. □

Example 11.7. Show that there is no field lying strictly between $\mathbb{Q}$ and $L = \mathbb{Q}[x]/(x^3 - 2)$.

Solution. The constant polynomials in L are identified with $\mathbb{Q}$. Suppose that K is a field such that $L \supseteq K \supseteq \mathbb{Q}$. Then $[L : \mathbb{Q}] = [L : K][K : \mathbb{Q}]$, by Theorem 11.6. But, by Theorem 11.5, $[L : \mathbb{Q}] = 3$, so $[L : K] = 1$, or $[K : \mathbb{Q}] = 1$.

If $[L : K] = 1$, then L is a vector space over K, and $\{1\}$, being linearly independent, is a basis. Hence $L = K$. If $[K : \mathbb{Q}] = 1$, then $K = \mathbb{Q}$. Hence there is no field lying strictly between L and $\mathbb{Q}$. □

Given a field extension K of F and an element $a \in K$, define $F(a)$ to be the intersection of all subfields of K that contain F and a. This is the *smallest* subfield of K containing F and a, and is called the field obtained by **adjoining** a to F.

For example, the smallest field containing $\mathbb{R}$ and i is the whole of the complex numbers, because this field must contain all elements of the form $a + ib$ where $a, b \in \mathbb{R}$. Hence $\mathbb{R}(i) = \mathbb{C}$.

In a similar way, the field obtained by adjoining $a_1, \ldots, a_n \in K$ to F is denoted by $F(a_1, \ldots, a_n)$ and is defined to be the smallest subfield of F containing $a_1, \ldots, a_n$ and F. It follows that $F(a_1, \ldots, a_n) = F(a_1, \ldots, a_{n-1})(a_n)$.

Example 11.8. $\mathbb{Q}(\sqrt{2})$ is equal to the subfield $F = \{a + b\sqrt{2} | a, b \in \mathbb{Q}\}$ of $\mathbb{R}$.

Solution. $\mathbb{Q}(\sqrt{2})$ must contain all rationals and $\sqrt{2}$. Hence $\mathbb{Q}(\sqrt{2})$ must contain all real numbers of the form $b\sqrt{2}$ for $b \in \mathbb{Q}$ and also $a + b\sqrt{2}$ for $a, b \in \mathbb{Q}$. Therefore, $F \subseteq \mathbb{Q}(\sqrt{2})$. But $\mathbb{Q}(\sqrt{2})$ is the smallest field containing $\mathbb{Q}$ and $\sqrt{2}$. Since F is another such field, $F \supseteq \mathbb{Q}(\sqrt{2})$ and so $F = \mathbb{Q}(\sqrt{2})$. □

If R is an integral domain and x is an indeterminate, then

$$R(x) = \left\{ \frac{a_0 + a_1 x + \cdots + a_n x^n}{b_0 + b_1 x + \cdots + b_m x^m} \middle| a_i, b_j \in R; \text{ not all the } b_j\text{'s are zero} \right\},$$

which is the field of **rational functions** in R. Any field containing R and x must contain the polynomial ring $R[x]$, and the smallest field containing $R[x]$ is its field of fractions $R(x)$.

ALGEBRAIC NUMBERS

If K is a field extension of F, the element $k \in K$ is called **algebraic** over F if there exist $a_0, a_1, \ldots, a_n \in F$, not all zero, such that

$$a_0 + a_1 k + \cdots + a_n k^n = 0.$$

In other words, k is the root of a nonzero polynomial in $F[x]$. Elements that are not algebraic over F are called **transcendental** over F.

For example, 5, $\sqrt{3}$, i, $\sqrt[n]{7} + 3$ are all algebraic over $\mathbb{Q}$ because they are roots of the polynomials $x - 5$, $x^2 - 3$, $x^2 + 1$, $(x - 3)^n - 7$, respectively.

Example 11.9. Find a polynomial in $\mathbb{Q}[x]$ with $\sqrt[3]{2} + \sqrt{5}$ as a root.

Solution. Let $x = \sqrt[3]{2} + \sqrt{5}$. We have to eliminate the square and cube roots from this equation. We have $x - \sqrt{5} = \sqrt[3]{2}$, so $(x - \sqrt{5})^3 = 2$ or $x^3 - 3\sqrt{5}x^2 + 15x - 5\sqrt{5} = 2$. Hence $x^3 + 15x - 2 = \sqrt{5}(3x^2 + 5)$, so $(x^3 + 15x - 2)^2 = 5(3x^2 + 5)^2$. Therefore, $\sqrt[3]{2} + \sqrt{5}$ is a root of $x^6 - 15x^4 - 4x^3 + 75x^2 - 60x - 121 = 0$. □

Not all real and complex numbers are algebraic over $\mathbb{Q}$. The numbers π and e can be proven to be transcendental over $\mathbb{Q}$ (see Stewart [35]). Since π is transcendental, we have

$$\mathbb{Q}(\pi) = \left\{ \frac{a_0 + a_1\pi + \cdots + a_n\pi^n}{b_0 + b_1\pi + \cdots + b_m\pi^m} \middle| a_i, b_j \in \mathbb{Q};\ \text{not all the } b_j\text{s are zero} \right\},$$

the field of rational functions in π with coefficients in $\mathbb{Q}$. $\mathbb{Q}(\pi)$ must contain all the powers of π and hence any polynomial in π with rational coefficients. Any nonzero element of $\mathbb{Q}(\pi)$ must have its inverse in $\mathbb{Q}(\pi)$; thus $\mathbb{Q}(\pi)$ contains the set of rational functions in π. The number $b_0 + b_1\pi + \cdots + b_m\pi^m$ is never zero unless $b_0 = b_1 = \cdots = b_m = 0$ because π is not the root of any polynomial with rational coefficients. This set of rational functions in π can be shown to be a subfield of $\mathbb{R}$.

Those readers acquainted with the theory of infinite sets can prove that the set of rational polynomials, $\mathbb{Q}[x]$, is countable. Since each polynomial has only a finite number of roots in $\mathbb{C}$, there are only a countable number of real or complex numbers algebraic over $\mathbb{Q}$. Hence there must be an uncountable number of real and complex numbers transcendental over $\mathbb{Q}$.

Example 11.10. Is $\cos(2\pi/5)$ algebraic or transcendental over $\mathbb{Q}$?

Solution. We know from De Moivre's theorem that

$$(\cos 2\pi/5 + i \sin 2\pi/5)^5 = \cos 2\pi + i \sin 2\pi = 1.$$

Taking real parts and writing $c = \cos 2\pi/5$ and $s = \sin 2\pi/5$, we have

$$c^5 - 10s^2c^3 + 5s^4c = 1.$$

Since $s^2 + c^2 = 1$, we have $c^5 - 10(1 - c^2)c^3 + 5(1 - c^2)^2c = 1$. That is, $16c^5 - 20c^3 + 5c - 1 = 0$ and hence $c = \cos 2\pi/5$ is algebraic over $\mathbb{Q}$. □

Theorem 11.11. Let α be algebraic over F and let $p(x)$ be an irreducible polynomial of degree n over F with α as a root. Then

$$F(\alpha) \cong F[x]/(p(x)),$$

and the elements of $F(\alpha)$ can be written uniquely in the form

$$c_0 + c_1\alpha + c_2\alpha^2 + \cdots + c_{n-1}\alpha^{n-1} \quad \text{where} \quad c_i \in F.$$

Proof. Define the ring morphism $f: F[x] \to F(\alpha)$ by $f(q(x)) = q(\alpha)$. The kernel of f is an ideal of $F[x]$. By Corollary 10.3, all ideals in $F[x]$ are principal; thus $\operatorname{Ker} f = (r(x))$ for some $r(x) \in F[x]$. Since $p(\alpha) = 0$, $p(x) \in \operatorname{Ker} f$, and so $r(x)|p(x)$. Since $p(x)$ is irreducible, $p(x) = kr(x)$ for some nonzero element k of F. Therefore, $\operatorname{Ker} f = (r(x)) = (p(x))$.

By the morphism theorem,

$$F[x]/(p(x)) \cong \operatorname{Im} f \subseteq F(\alpha).$$

Now, by Theorem 10.17, $F[x]/(p(x))$ is a field; thus $\operatorname{Im} f$ is a subfield of $F(\alpha)$ that contains F and α. Since $\operatorname{Im} f$ cannot be a smaller field than $F(\alpha)$, it follows that $\operatorname{Im} f = F(\alpha)$ and $F[x]/(p(x)) \cong F(\alpha)$.

The unique form for the elements of $F(\alpha)$ follows from the isomorphism above and Theorem 10.8. □

Corollary 11.12. If α is a root of the polynomial $p(x)$ of degree n, irreducible over F, then $[F(\alpha): F] = n$.

Proof. By Theorems 11.11 and 11.5, $[F(\alpha) : F] = [F[x]/(p(x)) : F] = n$. □

For example, $\mathbb{Q}(\sqrt{2}) \cong \mathbb{Q}[x]/(x^2 - 2)$ and $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$. Also, $\mathbb{Q}(\sqrt[4]{7}i) \cong \mathbb{Q}[x]/(x^4 - 7)$ and $[\mathbb{Q}(\sqrt[4]{7}i) : \mathbb{Q}] = 4$ because $\sqrt[4]{7}\, i$ is a root of $x^4 - 7$, which is irreducible over $\mathbb{Q}$, by Eisenstein's criterion (Theorem 9.30).

Lemma 11.13. Let $p(x)$ be an irreducible polynomial over the field F. Then F has a finite extension field K in which $p(x)$ has a root.

Proof. Let $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ and denote the ideal $(p(x))$ by P. By Theorem 11.5, $K = F[x]/P$ is a field extension of F of degree n whose elements are cosets of the form $P + f(x)$. The element $P + x \in K$ is a root of $p(x)$ because

$$\begin{aligned} a_0 + a_1(P + x) + a_2(P + x)^2 + \cdots + a_n(P + x)^n \\ &= a_0 + (P + a_1x) + (P + a_2x^2) + \cdots + (P + a_nx^n) \\ &= P + (a_0 + a_1x + a_2x^2 + \cdots + a_nx^n) = P + p(x) \\ &= P + 0, \end{aligned}$$

and this is the zero element of the field K. □

Theorem 11.14. If $f(x)$ is any polynomial over the field F, there is an extension field K of F over which $f(x)$ splits into linear factors.

Proof. We prove this by induction on the degree of $f(x)$. If $\deg f(x) \leqslant 1$, there is nothing to prove.

Suppose that the result is true for polynomials of degree $n - 1$. If $f(x)$ has degree n, we can factor $f(x)$ as $p(x)q(x)$, where $p(x)$ is irreducible over F. By Lemma 11.13, F has a finite extension K' in which $p(x)$ has a root, say α. Hence, by the factor theorem,

$$f(x) = (x - \alpha)g(x) \quad \text{where} \quad g(x) \text{ is of degree } n - 1 \text{ in } K'[x].$$

By the induction hypothesis, the field K' has a finite extension, K, over which $g(x)$ splits into linear factors. Hence $f(x)$ also splits into linear factors over K and, by Theorem 11.6, K is a finite extension of F. □

Let us now look at the development of the complex numbers from the real numbers. The reason for constructing the complex numbers is that certain equations, such as $x^2 + 1 = 0$, have no solution in $\mathbb{R}$. Since $x^2 + 1$ is a quadratic polynomial in $\mathbb{R}[x]$ without roots, it is irreducible over $\mathbb{R}$. In the above manner, we can extend the real field to

$$\mathbb{R}[x]/(x^2 + 1) = \{a + bx | a, b \in \mathbb{R}\}.$$

In this field extension

$$(0 + 1x)^2 = -1 \quad \text{since} \quad x^2 \equiv -1 \mod(x^2 + 1).$$

Denote the element $0 + 1x$ by i, so that $i^2 = -1$ and i is a root of the equation $x^2 + 1 = 0$ in this extension field. The field of complex numbers, $\mathbb{C}$, is defined to be $\mathbb{R}(i)$ and, by Theorem 11.11, there is an isomorphism

$$\psi: \mathbb{R}[x]/(x^2 + 1) \to \mathbb{R}(i)$$

defined by $\psi(a + bx) = a + bi$. Since

$$(a + bx) + (c + dx) \equiv (a + c) + (b + d)x \mod(x^2 + 1)$$

and

$$\begin{aligned}(a + bx)(c + dx) &\equiv ac + (ad + bc)x + bdx^2 \mod(x^2 + 1)\\ &\equiv (ac - bd) + (ad + bc)x \mod(x^2 + 1),\end{aligned}$$

addition and multiplication in $\mathbb{C} = \mathbb{R}(i)$ are defined in the standard way by

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

and

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i.$$

Example 11.15. Find $[\mathbb{Q}(\cos 2\pi/5) : \mathbb{Q}]$.

Solution. We know from Example 11.10 that $\cos 2\pi/5$ is algebraic over $\mathbb{Q}$ and is a root of the polynomial $16x^5 - 20x^3 + 5x - 1$. Using the same methods, we can show that $\cos 2k\pi/5$ is also a root of this equation for each $k \in \mathbb{Z}$. Hence we see from Figure 11.1 that its roots are $1, \cos 2\pi/5 = \cos 8\pi/5$, and $\cos 4\pi/5 = \cos 6\pi/5$. Therefore, $(x - 1)$ is a factor of the polynomial and

$$\begin{aligned}16x^5 - 20x^3 + 5x - 1 &= (x - 1)(16x^4 + 16x^3 - 4x^2 - 4x + 1)\\ &= (x - 1)(4x^2 + 2x - 1)^2.\end{aligned}$$

It follows that $\cos 2\pi/5$ and $\cos 4\pi/5$ are roots of the quadratic $4x^2 + 2x - 1$ so by the quadratic formula, these roots are $(-1 \pm \sqrt{5})/4$. Since $\cos 2\pi/5$ is positive,

$$\cos 2\pi/5 = (\sqrt{5} - 1)/4 \quad \text{and} \quad \cos 4\pi/5 = (-\sqrt{5} - 1)/4.$$

Therefore, $\mathbb{Q}(\cos 2\pi/5) \cong \mathbb{Q}[x]/(4x^2 + 2x - 1)$ because $4x^2 + 2x - 1$ is irreducible over $\mathbb{Q}$. By Corollary 11.12, $[\mathbb{Q}(\cos 2\pi/5) : \mathbb{Q}] = 2$. □

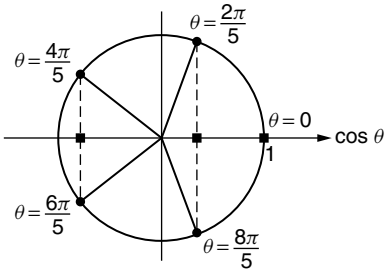


Figure 11.1. Values of $\cos(2k\pi/5)$.

Proposition 11.16. If $[K : F] = 2$, where $F \supseteq \mathbb{Q}$, then $K = F(\sqrt{\gamma})$ for some $\gamma \in F$.

Proof. K is a vector space of dimension 2 over F. Extend $\{1\}$ to a basis $\{1, \alpha\}$ for K over F, so that $K = \{a\alpha + b | a, b \in F\}$.

Now K is a field, so $\alpha^2 \in K$ and $\alpha^2 = a\alpha + b$ for some $a, b \in F$. Hence $(\alpha - (a/2))^2 = b + (a^2/4)$. Put $\beta = \alpha - (a/2)$. Then $\{1, \beta\}$ is also a basis for K over F, and $K = F(\beta)$ where $\beta^2 = b + (a^2/4) = \gamma \in F$. Hence $K = F(\sqrt{\gamma})$. □

Proposition 11.17. If F is an extension field of $\mathbb{R}$ of finite degree, then F is isomorphic to $\mathbb{R}$ or $\mathbb{C}$.

Proof. Let $[F : \mathbb{R}] = n$. If $n = 1$, then F is equal to $\mathbb{R}$. Otherwise, $n > 1$ and F contains some element α not in $\mathbb{R}$. Now $\{1, \alpha, \alpha^2, \ldots, \alpha^n\}$ is a linearly dependent set of elements of F over $\mathbb{R}$, because it contains more than n elements; hence there exist real numbers $\lambda_0, \lambda_1, \ldots, \lambda_n$, not all zero, such that

$$\lambda_0 + \lambda_1\alpha + \cdots + \lambda_n\alpha^n = 0.$$

The element α is therefore algebraic over $\mathbb{R}$ so since the only irreducible polynomials over $\mathbb{R}$ have degree 1 or 2, α must satisfy a linear or quadratic equation over $\mathbb{R}$. If it satisfies a linear equation, then $\alpha \in \mathbb{R}$, contrary to our hypothesis. Therefore, $[\mathbb{R}(\alpha) : \mathbb{R}] = 2$, and, by Proposition 11.16, $\mathbb{R}(\alpha) = \mathbb{R}(\sqrt{\gamma})$. In this case γ must be negative because $\mathbb{R}$ contains all positive square roots; hence $\sqrt{\gamma} = ir$, where $r \in \mathbb{R}$ and $\mathbb{R}(\alpha) = \mathbb{R}(i) \cong \mathbb{C}$.

Therefore, the field F contains a subfield $C = \mathbb{R}(\alpha)$ isomorphic to the complex numbers and $[F : C] = [F : \mathbb{R}]/2$, which is finite. By an argument similar to the above, any element of C is the root of an irreducible polynomial over C. However, the only irreducible polynomials over C are the linear polynomials, and all their roots lie in C. Hence $F = C$ is isomorphic to $\mathbb{C}$. □

Example 11.18. $[\mathbb{R} : \mathbb{Q}]$ is infinite.

Solution. The real number $\sqrt[n]{2}$ is a root of the polynomial $x^n - 2$, which, by Eisenstein's criterion, is irreducible over $\mathbb{Q}$. If $[\mathbb{R} : \mathbb{Q}]$ were finite, we could use Theorem 11.6 and Corollary 11.12 to show that

$$[\mathbb{R} : \mathbb{Q}] = [\mathbb{R} : \mathbb{Q}(\sqrt[n]{2})][\mathbb{Q}(\sqrt[n]{2}) : \mathbb{Q}] = [\mathbb{R} : \mathbb{Q}(\sqrt[n]{2})]n.$$

This is a contradiction, because no finite integer is divisible by every integer n. Hence $[\mathbb{R} : \mathbb{Q}]$ must be infinite. □

GALOIS FIELDS

In this section we investigate the structure of finite fields; these fields are called Galois fields in honor of the mathematician Évariste Galois (1811–1832).

We show that the element 1 in any finite field generates a subfield isomorphic to $\mathbb{Z}_p$, for some prime p called the characteristic of the field. Hence a finite field is some finite extension of the field $\mathbb{Z}_p$ and so must contain p^m elements, for some integer m.

The characteristic can be defined for any ring, and we give the general definition here, even though we are mainly interested in its application to fields.

For any ring R, define the ring morphism $f:\mathbb{Z} \to R$ by $f(n) = n \cdot 1_R$ where 1_R is the identity of R. The kernel of f is an ideal of the principal ideal ring $\mathbb{Z}$; hence $\text{Ker} f = (q)$ for some $q \geqslant 0$. The generator $q \geqslant 0$ of $\text{Ker} f$ is called the **characteristic** of the ring R. If $a \in R$ then $qa = q(1_R a) = (q1_R)a = 0$, Hence if $q > 0$ the characteristic of R is the least integer $q > 0$ for which $qa = 0$, for all $a \in R$. If no such number exists, the characteristic of R is zero. For example, the characteristic of $\mathbb{Z}$ is 0, and the characteristic of $\mathbb{Z}_n$ is n.

Proposition 11.19. The characteristic of an integral domain is either zero or prime.

Proof. Let q be the characteristic of an integral domain D. By applying the morphism theorem to $f:\mathbb{Z} \to D$, defined by $f(1) = 1$, we see that

$$f(\mathbb{Z}) \cong \begin{cases} \mathbb{Z}_q & \text{if } q \neq 0 \\ \mathbb{Z} & \text{if } q = 0. \end{cases}$$

But $f(\mathbb{Z})$ is a subring of an integral domain; therefore, it has no zero divisors, and, by Theorem 8.11, q must be zero or prime. □

The characteristic of the field $\mathbb{Z}_p$ is p, while $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$ have zero characteristic.

Proposition 11.20. If the field F has prime characteristic p, then F contains a subfield isomorphic to $\mathbb{Z}_p$. If the field F has zero characteristic, then F contains a subfield isomorphic to the rational numbers, $\mathbb{Q}$.

Proof. From the proof of Proposition 11.19 we see that F contains the subring $f(\mathbb{Z})$, which is isomorphic to $\mathbb{Z}_p$ if F has prime characteristic p. If the characteristic of F is zero, $f:\mathbb{Z} \to f(\mathbb{Z})$ is an isomorphism. We show that F contains the field of fractions of $f(\mathbb{Z})$ and that this is isomorphic to $\mathbb{Q}$.

Let $Q = \{xy^{-1} \in F | x, y \in f(\mathbb{Z})\}$, a subring of F. Define the function

$$\tilde{f}:\mathbb{Q} \to Q$$

by $\tilde{f}(a/b) = f(a) \cdot f(b)^{-1}$. Since rational numbers are defined as equivalence classes, we have to check that $\tilde{f}$ is well defined. We can show that $\tilde{f}(a/b) = \tilde{f}(c/d)$ if $a/b = c/d$. Furthermore, it can be verified that $\tilde{f}$ is a ring isomorphism. Hence Q is isomorphic to $\mathbb{Q}$. □

Corollary 11.21. The characteristic of a finite field is nonzero.

Theorem 11.22. If F is a finite field, it has p^m elements for some prime p and some integer m.

Proof. By the previous results, F has characteristic p, for some prime p, and contains a subfield isomorphic to $\mathbb{Z}_p$. We identify this subfield with $\mathbb{Z}_p$ so that F is a field extension of $\mathbb{Z}_p$. The degree of this extension must be finite because F is finite. Let $[F : \mathbb{Z}_p] = m$ and let $\{f_1, \ldots, f_m\}$ be a basis of F over $\mathbb{Z}_p$, so that

$$F = \{\lambda_1 f_1 + \cdots + \lambda_m f_m | \lambda_i \in \mathbb{Z}_p\}.$$

There are p choices for each λ_i; therefore, F contains p^m elements. □

A finite field with p^m elements is called a **Galois field** of order p^m and is denoted by $GF(p^m)$. It can be shown that for a given prime p and positive integer m, a Galois field $GF(p^m)$ exists and that all fields of order p^m are isomorphic. See Stewart [35] or Nicholson [11] for a proof of these facts. For $m = 1$, the integers modulo p, $\mathbb{Z}_p$, is a Galois field of order p.

From Theorem 11.22 it follows that $GF(p^m)$ is a field extension of $\mathbb{Z}_p$ of degree m. Each finite field $GF(p^m)$ can be constructed by finding a polynomial $q(x)$ of degree m, irreducible in $\mathbb{Z}_p[x]$, and defining

$$GF(p^m) = \mathbb{Z}_p[x]/(q(x)).$$

By Lemma 11.13 and Corollary 11.12, there is an element α in $GF(p^m)$, such that $q(\alpha) = 0$, and $GF(p^m) = \mathbb{Z}_p(\alpha)$, the field obtained by adjoining α to $\mathbb{Z}_p$.

For example, $GF(4) = \mathbb{Z}_2[x]/(x^2 + x + 1) = \mathbb{Z}_2(\alpha) = \{0, 1, \alpha, \alpha + 1\}$, where $\alpha^2 + \alpha + 1 = 0$. Rewriting Table 10.2, we obtain Table 11.1 for $GF(4)$.

Example 11.23. Construct a field $GF(125)$.

Solution. Since $125 = 5^3$, we can construct such a field if we can find an irreducible polynomial of degree 3 over $\mathbb{Z}_5$.

A reducible polynomial of degree 3 must have a linear factor. Therefore, by the factor theorem, $p(x) = x^3 + ax^2 + bx + c$ is irreducible in $\mathbb{Z}_5[x]$ if and only if $p(n) \neq 0$ for $n = 0, 1, 2, 3, 4$ in $\mathbb{Z}_5$.

TABLE 11.1. Galois Field *GF* (4)

$+$	0	1	α	$\alpha+1$
0	0	1	α	$\alpha+1$
1	1	0	$\alpha+1$	α
α	α	$\alpha+1$	0	1
$\alpha+1$	$\alpha+1$	α	1	0

$\cdot$	0	1	α	$\alpha+1$
0	0	0	0	0
1	0	1	α	$\alpha+1$
α	0	α	$\alpha+1$	1
$\alpha+1$	0	$\alpha+1$	1	α

By trial and error, we find that the polynomial $p(x) = x^3 + x + 1$ is irreducible because $p(0) = 1$, $p(1) = 3$, $p(2) = 11 = 1$, $p(3) = 31 = 1$, and $p(4) = p(-1) = -1 = 4$ in $\mathbb{Z}_5$. Hence

$$GF(125) = \mathbb{Z}_5[x]/(x^3 + x + 1). \qquad \square$$

Note that $(x^3 + x + 1)$ is not the only irreducible polynomial of degree 3 over $\mathbb{Z}_5$. For example, $(x^3 + x^2 + 1)$ is also irreducible. But $\mathbb{Z}_5[x]/(x^3 + x + 1)$ is isomorphic to $\mathbb{Z}_5[x]/(x^3 + x^2 + 1)$.

PRIMITIVE ELEMENTS

The elements of a Galois field $GF(p^m)$ can be written as

$$\left\{a_0 + a_1\alpha + \cdots + a_{m-1}\alpha^{m-1} | a_i \in \mathbb{Z}_p\right\}$$

where α is a root of a polynomial $q(x)$ of degree m irreducible over $\mathbb{Z}_p$. Addition is easily performed using this representation, because it is simply addition of polynomials in $\mathbb{Z}_p[\alpha]$. However, multiplication is more complicated and requires repeated use of the relation $q(\alpha) = 0$. We show that by judicious choice of α, the elements of $GF(p^m)$ can be written as

$$\left\{0, 1, \alpha, \alpha^2, \alpha^3, \ldots, \alpha^{p^m-2}\right\} \quad \text{where} \quad \alpha^{p^m-1} = 1.$$

This element α is called a *primitive element* of $GF(p^m)$, and multiplication is easily calculated using powers of α; however, addition is much harder to perform using this representation.

For example, in $GF(4) = \mathbb{Z}_2(\alpha) = \{0, 1, \alpha, \alpha^2\}$ where $\alpha + 1 = \alpha^2$ and $\alpha^3 = 1$, and the tables are given in Table 11.2.

If F is any field and $F^* = F - \{0\}$, we know that $(F^*, \cdot)$ is an abelian group under multiplication. We now show that the nonzero elements of a finite field form a cyclic group under multiplication; the generators of this cyclic group are the primitive elements of the field. To prove this theorem, we need some preliminary results about the orders of elements in an abelian group.

TABLE 11.2. Galois Field *GF* (4) in Terms of a Primitive Element

$+$	0	1	α	α^2
0	0	1	α	α^2
1	1	0	α^2	α
α	α	α^2	0	1
α^2	α^2	α	1	0

$\cdot$	0	1	α	α^2
0	0	0	0	0
1	0	1	α	α^2
α	0	α	α^2	1
α^2	0	α^2	1	α

Lemma 11.24. If g and h are elements of an abelian group of orders a and b, respectively, there exists an element of order $\text{lcm}(a, b)$.

Proof. Let $a = p_1^{a_1} p_2^{a_2} \cdots p_s^{a_s}$ and $b = p_1^{b_1} p_2^{b_2} \cdots p_s^{b_s}$, where the p_i are distinct primes and $a_i \geqslant 0, b_i \geqslant 0$ for each i. For each i define

$$x_i = \begin{cases} a_i & \text{if } a_i \geqslant b_i \\ 0 & \text{if } a_i < b_i \end{cases} \quad \text{and} \quad y_i = \begin{cases} 0 & \text{if } a_i \geqslant b_i \\ b_i & \text{if } a_i < b_i \end{cases}.$$

If $x = p_1^{x_1} p_2^{x_2} \cdots p_s^{x_s}$ and $y = p_1^{y_1} p_2^{y_2} \cdots p_s^{y_s}$, then $x|a$ and $y|b$, so $g^{a/x}$ has order x and $h^{b/y}$ has order y. Moreover, $\gcd(x, y) = 1$, so $g^{a/x} h^{b/y}$ has order xy by Lemma 4.36. But $xy = \text{lcm}(a, b)$ by Theorem 15 of Appendix 2 because $x_i + y_i = \max(a_i, b_i)$ for each i. □

Lemma 11.25. If the maximum order of the elements of an abelian group G is r, then $x^r = e$ for all $x \in G$. □

Proof. Let $g \in G$ be an element of maximal order r. If h is an element of order t, there is an element of order $\text{lcm}(r, t)$ by Lemma 11.24. Since $\text{lcm}(r, t) \leqslant r$, t divides r. Therefore, $h^r = e$. □

Theorem 11.26. Let $GF(q)^*$ be the set of nonzero elements in the Galois field $GF(q)$. Then $(GF(q)^*, \cdot)$ is a cyclic group of order $q - 1$.

Proof. Let r be the maximal order of elements of $(GF(q)^*, \cdot)$. Then, by Lemma 11.25,

$$x^r - 1 = 0 \quad \text{for all } x \in GF(q)^*.$$

Hence every nonzero element of the Galois field $GF(q)$ is a root of the polynomial $x^r - 1$ and, by Theorem 9.6, a polynomial of degree r can have at most r roots over any field; therefore, $r \geqslant q - 1$. But, by Lagrange's theorem, $r|(q - 1)$; it follows that $r = q - 1$.

$(GF(q)^*, \cdot)$ is therefore a group of order $q - 1$ containing an element of order $q - 1$ and hence must be cyclic. □

A generator of the cyclic group $(GF(q)^*, \cdot)$ is called a **primitive element** of $GF(q)$. For example, in $GF(4) = \mathbb{Z}_2(\alpha)$, the multiplicative group of nonzero elements, $GF(4)^*$, is a cyclic group of order 3, and both nonidentity elements α and $\alpha + 1$ are primitive elements.

If α is a primitive element in the Galois field $GF(q)$, where q is the power of a prime p, then $GF(q)$ is the field extension $\mathbb{Z}_p(\alpha)$ and

$$GF(q)^* = \{1, \alpha, \alpha^2, \ldots, \alpha^{q-2}\}.$$

Hence

$$GF(q) = \{0, 1, \alpha, \alpha^2, \ldots, \alpha^{q-2}\}.$$

Example 11.27. Find all the primitive elements in $GF(9) = \mathbb{Z}_3(\alpha)$, where $\alpha^2 + 1 = 0$.

Solution. Since $x^2 + 1$ is irreducible over $\mathbb{Z}_3$, we have

$$GF(9) = \mathbb{Z}_3[x]/(x^2 + 1) = \{a + bx | a, b \in \mathbb{Z}_3\}.$$

The nonzero elements form a cyclic group $GF(9)^*$ of order 8; hence the multiplicative order of each element is either 1, 2, 4, or 8.

In calculating the powers of each element, we use the relationship $\alpha^2 = -1 = 2$. From Table 11.3, we see that $1 + \alpha$, $2 + \alpha$, $1 + 2\alpha$, and $2 + 2\alpha$ are the primitive elements of $GF(9)$. □

Proposition 11.28.

(i) $z^{q-1} = 1$ for all elements $z \in GF(q)^*$.
(ii) $z^q = z$ for all elements $z \in GF(q)$.
(iii) If $GF(q) = \{\alpha_1, \alpha_2, \ldots, \alpha_q\}$, then $z^q - z$ factors over $GF(q)$ as

$$(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_q).$$

Proof. We have already shown that (i) is implied by Lemma 11.25. Part (ii) follows immediately because 0 is the only element of $GF(q)$ that is not in $GF(q)^*$. The polynomial $z^q - z$, of degree q, can have at most q roots over any field. By (ii), *all* elements of $GF(q)$ are roots over $GF(q)$; hence $z^q - z$ factors into q distinct linear factors over $GF(q)$. □

For example, in $GF(4) = \mathbb{Z}_2[x]/(x^2 + x + 1) = \{0, 1, \alpha, \alpha + 1\}$, and we have

$$\begin{aligned}(z + 0)(z + 1)(z + \alpha)(z + \alpha + 1) &= (z^2 + z)(z^2 + z + \alpha^2 + \alpha) \\ &= (z^2 + z)(z^2 + z + 1) \\ &= z^4 + z = z^4 - z.\end{aligned}$$

TABLE 11.3. Nonzero Elements of *GF* (9)

Element x	x^2	x^4	x^8	Order	Primitive
1	1	1	1	1	No
2	1	1	1	2	No
α	2	1	1	4	No
$1 + \alpha$	2α	2	1	8	Yes
$2 + \alpha$	α	2	1	8	Yes
2α	2	1	1	4	No
$1 + 2\alpha$	α	2	1	8	Yes
$2 + 2\alpha$	2α	2	1	8	Yes

An irreducible polynomial $g(x)$, of degree m over $\mathbb{Z}_p$, is called a **primitive polynomial** if $g(x)|(x^k - 1)$ for $k = p^m - 1$ and for no smaller k.

Proposition 11.29. The irreducible polynomial $g(x) \in \mathbb{Z}_p[x]$ is primitive if and only if x is a primitive element in $\mathbb{Z}_p[x]/(g(x)) = GF(p^m)$.

Proof. The following statements are equivalent:

(i) x is a primitive element in $GF(p^m) = \mathbb{Z}_p[x]/(g(x))$.
(ii) $x^k = 1$ in $GF(p^m)$ for $k = p^m - 1$ and for no smaller k.
(iii) $x^k - 1 \equiv 0 \bmod g(x)$ for $k = p^m - 1$ and for no smaller k.
(iv) $g(x)|(x^k - 1)$ for $k = p^m - 1$ and for no smaller k. □

For example, $x^2 + x + 1$ is primitive in $\mathbb{Z}_2[x]$. From Example 11.27, we see that $x^2 + 1$ is not primitive in $\mathbb{Z}_3[x]$. However, $1 + \alpha$ and $1 + 2\alpha = 1 - \alpha$ are primitive elements, and they are roots of the polynomial

$$(x - 1 - \alpha)(x - 1 + \alpha) = (x - 1)^2 - \alpha^2 = x^2 + x + 2 \in \mathbb{Z}_3[x].$$

Hence $x^2 + x + 2$ is a primitive polynomial in $\mathbb{Z}_3[x]$. Also, $x^2 + 2x + 2$ is another primitive polynomial in $\mathbb{Z}_3[x]$ with roots $2 + \alpha$ and $2 + 2\alpha = 2 - \alpha$.

Example 11.30. Let α be a root of the primitive polynomial $x^4 + x + 1 \in \mathbb{Z}_2[x]$. Show how the nonzero elements of $GF(16) = \mathbb{Z}_2(\alpha)$ can be represented by the powers of α.

Solution. The representation is given in Table 11.4. □

Arithmetic in $GF(16)$ can very easily be performed using Table 11.4. Addition is performed by representing elements as polynomials in α of degree less than 4, whereas multiplication is performed using the representation of nonzero elements as powers of α. For example,

$$\begin{aligned}
\frac{1 + \alpha + \alpha^3}{1 + \alpha^2 + \alpha^3} + \alpha + \alpha^2 &= \frac{\alpha^7}{\alpha^{13}} + \alpha + \alpha^2 \\
&= \alpha^{-6} + \alpha + \alpha^2 \\
&= \alpha^9 + \alpha + \alpha^2 \qquad \text{since } \alpha^{15} = 1 \\
&= \alpha + \alpha^3 + \alpha + \alpha^2 \\
&= \alpha^2 + \alpha^3.
\end{aligned}$$

The concept of primitive polynomials is useful in designing **feedback shift registers** with a long cycle length. Consider the circuit in Figure 11.2, in which the square boxes are delays of one unit of time, and the circle with a cross inside represents a modulo 2 adder.

TABLE 11.4. Representation of *GF*(16)

Element	α^0	α^1	α^2	α^3
$0 = 0$	0	0	0	0
$\alpha^0 = 1$	1	0	0	0
$\alpha^1 = \alpha$	0	1	0	0
$\alpha^2 = \alpha^2$	0	0	1	0
$\alpha^3 = \alpha^3$	0	0	0	1
$\alpha^4 = 1 + \alpha$	1	1	0	0
$\alpha^5 = \alpha + \alpha^2$	0	1	1	0
$\alpha^6 = \alpha^2 + \alpha^3$	0	0	1	1
$\alpha^7 = 1 + \alpha + \alpha^3$	1	1	0	1
$\alpha^8 = 1 + \alpha^2$	1	0	1	0
$\alpha^9 = \alpha + \alpha^3$	0	1	0	1
$\alpha^{10} = 1 + \alpha + \alpha^2$	1	1	1	0
$\alpha^{11} = \alpha + \alpha^2 + \alpha^3$	0	1	1	1
$\alpha^{12} = 1 + \alpha + \alpha^2 + \alpha^3$	1	1	1	1
$\alpha^{13} = 1 + \alpha^2 + \alpha^3$	1	0	1	1
$\alpha^{14} = 1 + \alpha^3$	1	0	0	1
$\alpha^{15} = 1$				

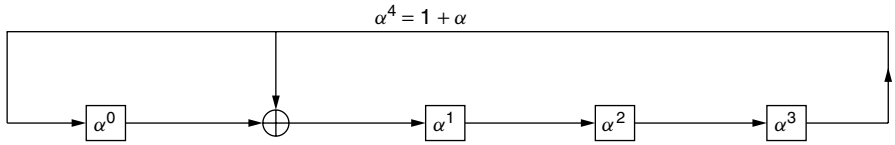


Figure 11.2. Feedback shift register.

If the delays are labeled by a representation of the elements of $GF(16)$, a single shift corresponds to multiplying the element of $GF(16)$ by α. Hence, if the contents of the delays are not all zero initially, this shift register will cycle through 15 different states before repeating itself. In general, it is possible to construct a shift register with n delay units that will cycle through $2^n - 1$ different states before repeating itself. The feedback connections have to be derived from a primitive polynomial of degree n over $\mathbb{Z}_2$. Such feedback shift registers are useful in designing error-correcting coders and decoders, random number generators, and radar transmitters. See Chapter 14 of this book and, Lidl and Niederreiter [34, Chap. 6], or Stone [22, Chap. 9].

EXERCISES

For Exercises 11.1 to 11.4, write out the addition and multiplication tables for the fields.

11.1. $GF(5)$.
11.2. $GF(7)$.
11.3. $GF(9)$.
11.4. $GF(8)$.

For Exercises 11.5 to 11.10, in each case find, if possible, an irreducible polynomial of degree n over F.

11.5. $n = 3$, $F = \mathbb{Z}_{11}$.
11.6. $n = 3$, $F = \mathbb{Q}$.
11.7. $n = 4$, $F = \mathbb{R}$.
11.8. $n = 3$, $F = GF(4)$.
11.9. $n = 2$, $F = \mathbb{Q}(i)$.
11.10. $n = 5$, $F = \mathbb{Z}_3$.

For Exercises 11.11 to 11.13, in each case, find a polynomial in $F[x]$ with r as a root.

11.11. $r = \sqrt{2} + \sqrt{6}$, $F = \mathbb{Q}$.
11.12. $r = \pi + ei$, $F = \mathbb{R}$.
11.13. $r = \sqrt[3]{3}/\sqrt{2}$, $F = \mathbb{Q}$.
11.14. Show that $\theta = 2k\pi/7$ satisfies the equation $\cos 4\theta - \cos 3\theta = 0$ for each integer k. Hence find an irreducible polynomial over $\mathbb{Q}$ with $\cos(2\pi/7)$ as a root.
11.15. Prove that the **algebraic numbers**

$$\mathbb{A} = \{x \in \mathbb{C} | x \text{ is algebraic over } \mathbb{Q}\}$$

form a subfield of $\mathbb{C}$.
11.16. Assuming the fundamental theorem of algebra, prove that every polynomial in $\mathbb{A}$ has a root in $\mathbb{A}$.

For Exercises 11.17 to 11.25, Calculate the degrees.

11.17. $[\mathbb{Q}(\sqrt[3]{7}) : \mathbb{Q}]$.
11.18. $[\mathbb{C} : \mathbb{Q}]$.
11.19. $[\mathbb{Q}(i, 3i) : \mathbb{Q}]$.
11.20. $[\mathbb{C} : \mathbb{R}(\sqrt{-7})]$.
11.21. $[\mathbb{Z}_3[x]/(x^2 + x + 2) : \mathbb{Z}_3]$.
11.22. $[\mathbb{Q}(i, \sqrt{2}) : \mathbb{Q}]$.
11.23. $[\mathbb{A} : \mathbb{Q}]$.
11.24. $[\mathbb{C} : \mathbb{A}]$.
11.25. $[\mathbb{Z}_3(t) : \mathbb{Z}_3]$, where $\mathbb{Z}_3(t)$ is the field of rational functions in t over $\mathbb{Z}_3$.
11.26. Prove that $x^2 - 2$ is irreducible over $\mathbb{Q}(\sqrt{3})$.

For Exercises 11.27 to 11.32, find the inverses of the elements in the given fields. Each field is a finite extension $F(\alpha)$. Express your answers in the form $a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}$, where $a_i \in F$ and $[F(\alpha) : F] = n$.

11.27. $1 + \sqrt[3]{2}$ in $\mathbb{Q}(\sqrt[3]{2})$.
11.28. $\sqrt[4]{5} + \sqrt{5}$ in $\mathbb{Q}(\sqrt[4]{5})$.
11.29. $5 + 6\omega$ in $\mathbb{Q}(\omega)$, where ω is a complex cube root of 1.

11.30. $2 - 3i$ in $\mathbb{Q}(i)$.

11.31. α in $GF(32) = \mathbb{Z}_2(\alpha)$, where $\alpha^5 + \alpha^2 + 1 = 0$.

11.32. α in $GF(27) = \mathbb{Z}_3(\alpha)$, where $\alpha^3 + \alpha^2 + 2 = 0$.

For Exercises 11.33 to 11.40, find the characteristic of the rings. Which of these are fields?

11.33. $\mathbb{Z}_2 \times \mathbb{Z}_2$.

11.34. $\mathbb{Z}_3 \times \mathbb{Z}_4$.

11.35. $GF(49)$.

11.36. $\mathbb{Z} \times \mathbb{Z}_2$.

11.37. $\mathbb{Q}(\sqrt[3]{7})$.

11.38. $M_2(\mathbb{Z}_5)$.

11.39. $\mathbb{Q} \times \mathbb{Z}_3$.

11.40. $GF(4)[x]$.

11.41. Let R be any ring and n a positive integer. Prove that $I_n = \{na | a \in R\}$ is an ideal of R and that the characteristic of R/I_n divides n.

11.42. Let M be a finite subgroup of the multiplicative group F^* of any infinite field F. Prove that M is cyclic, and give an example to show that F^* need not be cyclic.

11.43. For what values of m is $(\mathbb{Z}_m^*, \cdot)$ cyclic? (This is a difficult problem; see Exercises 4.55 to 4.62 for other results on $\mathbb{Z}_m^*$.)

11.44. Let $GF(4) = \mathbb{Z}_2(\alpha)$, where $\alpha^2 + \alpha + 1 = 0$. Find an irreducible quadratic in $GF(4)[x]$. If β is the root of such a polynomial, show that $GF(4)(\beta)$ is a Galois field of order 16.

11.45. **(a)** Show that there are $(p^2 - p)/2$ monic irreducible polynomials of degree 2 over $GF(p)$. (A polynomial is **monic** if the coefficient of the highest power of the variable is 1.)

(b) Prove that there is a field with p^2 elements for every prime p.

11.46. **(a)** How many monic irreducible polynomials of degree 3 are there over $GF(p)$?

(b) Prove that there is a field with p^3 elements for every prime p.

11.47. Find an element α such that $\mathbb{Q}(\sqrt{2}, \sqrt{-3}) = \mathbb{Q}(\alpha)$.

11.48. Find all primitive elements in $GF(16) = \mathbb{Z}_2(\alpha)$, where $\alpha^4 + \alpha + 1 = 0$.

11.49. Find all the primitive elements in $GF(32)$.

For Exercises 11.50 and 11.51, find a primitive polynomial of degree n over the field F.

11.50. $n = 2$, $F = \mathbb{Z}_5$.

11.51. $n = 3$, $F = \mathbb{Z}_2$.

11.52. Let $g(x)$ be a polynomial of degree m over $\mathbb{Z}_p$. If $g(x) | (x^k - 1)$ for $k = p^m - 1$ and for no smaller k, show that $g(x)$ is irreducible over $\mathbb{Z}_p$.

11.53. Prove that $x^8 + x \in \mathbb{Z}_2[x]$ will split into linear factors over $GF(8)$ but not over any smaller field.

11.54. Let $f(x) = 2x^3 + 5x^2 + 7x + 6 \in \mathbb{Q}[x]$. Find a field, smaller than the complex numbers, in which $f(x)$ splits into linear factors.

11.55. If α and β are roots of $x^3 + x + 1$ and $x^3 + x^2 + 1 \in \mathbb{Z}_2[x]$, respectively, prove that the Galois fields $\mathbb{Z}_2(\alpha)$ and $\mathbb{Z}_2(\beta)$ are isomorphic.

11.56. **(a)** If $p(x) \in \mathbb{Z}_2[x]$, prove that $[p(x)]^2 = p(x^2)$.
(b) If β is a root of $p(x) \in \mathbb{Z}_2[x]$, prove that β^{2^l} is a root for all $l \in \mathbb{N}$.
(c) Let $GF(16) = \mathbb{Z}_2(\alpha)$ where $\alpha^4 + \alpha + 1 = 0$. Find an irreducible polynomial in $\mathbb{Z}_2[x]$ with α^3 as a root.

For Exercises 11.57 and 11.58, solve the simultaneous linear equations in $GF(4) = \mathbb{Z}_2(\alpha)$.

11.57. $\alpha x + (\alpha + 1)y = \alpha + 1$
$x + \alpha y = 1$.

11.58. $(\alpha + 1)x + y = \alpha$
$x + (\alpha + 1)y = \alpha + 1$.

11.59. Solve the quadratic equation $\alpha x^2 + (1 + \alpha)x + 1 = 0$ over the field $GF(4) = \mathbb{Z}_2(\alpha)$.

11.60. Let R be any commutative ring of characteristic p, where p is a prime.
(a) Show that $(a + b)^p = a^p + b^p$ for all a, b in R. [*Hint*: If $\binom{p}{r}$ denotes the binomial coefficient, show that $\binom{p}{r} = \frac{p}{r}\binom{p-1}{r-1}$ whenever $1 \leqslant r \leqslant p$.]
(b) If $\sigma : R \to R$ is defined by $\sigma(a) = a^p$ for all $a \in R$, show that σ is a ring morphism.
(c) If $R = GF(p^n)$ show that σ is an isomorphism of R (called the **Frobenius automorphism**).

11.61. If F is a field and F^* is cyclic, show that F is finite.

11.62. Design a feedback shift register using six delays that has a cycle length of 63.

11.63. What is the cycle length of the feedback shift register in Figure 11.3?

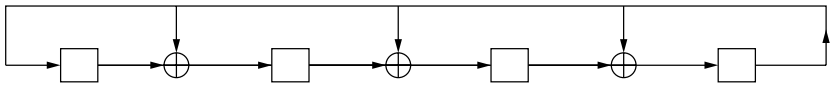


Figure 11.3

11.64. Design a feedback shift register that has a cycle length of 21.

11.65. Describe the output sequence of the feedback shift register in Figure 11.4 when the registers initially contain the elements shown.

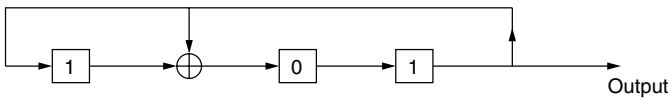


Figure 11.4

11.66. If a feedback shift register with n delays has a cycle length of $2^n - 1$, show that the feedback connections must be derived from a primitive irreducible polynomial of degree n over $\mathbb{Z}_2$.

12

LATIN SQUARES

Latin squares first arose with parlor games such as the problem of arranging the jacks, queens, kings, and aces of a pack of cards in a 4×4 array so that each row and each column contains one card from each suit and one card from each rank. In 1779, Leonard Euler posed the following famous problem of the 36 officers from six ranks and six regiments. He claimed that it was impossible to arrange these officers on parade in a 6×6 square so that each row and each column contains one officer from each rank and one from each regiment.

Recently, statisticians have found latin squares useful in designing experiments, and mathematicians have found close connections between latin squares and finite geometries.

LATIN SQUARES

Let S be a set with n elements. Then a **latin square** $L = (l_{ij})$, of order n based on S, is an $n \times n$ array of the elements of S such that each element appears exactly once in each row and once in each column.

For example, Table 12.1 illustrates a latin square of order 3 based on $\{a, b, c\}$.

Theorem 12.1. The table for any finite group $(G, +)$ of order n is a latin square of order n based on G.

Proof. We write the operation in G as addition, even though the result still holds if G is not commutative.

Suppose that two elements in one row are equal. Then $x_i + x_j = x_i + x_k$ for some $x_i, x_j, x_k \in G$. Since G is a group, x_i has an inverse $(-x_i)$ such that $(-x_i) + x_i = 0$. Hence $(-x_i) + (x_i + x_j) = (-x_i) + (x_i + x_k)$, and since the operation is associative, we have $x_j = x_k$. Therefore, an element cannot appear twice in the same row. Similarly, an element cannot appear twice in the same column, and the table is a latin square. □

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

TABLE 12.1. Latin Square

a	b	c
c	a	b
b	c	a

TABLE 12.2. Latin Squares of Order 4

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

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

Given any latin square, we can permute the rows among themselves and also the columns among themselves and we still have a latin square. For example, the addition table for $\mathbb{Z}_2 \times \mathbb{Z}_2$ is a latin square of order 4. If we interchange the first and third columns and replace (0, 0) by a, (0, 1) by b, (1, 0) by c, and (1, 1) by d, we obtain another latin square of order 4 based on $\{a, b, c, d\}$. These are illustrated in Table 12.2.

Latin squares are useful in designing statistical experiments because they can show how an experiment can be arranged so as to reduce the errors without making the experiment too large or too complicated. See Laywine and Mullen [40] for more complete details.

Suppose that you wanted to compare the yields of three varieties of hybrid corn. You have a rectangular test plot, but you are not sure that the fertility of the soil is the same everywhere. You could divide up the land into nine rectangular regions and plant the three varieties, *a, b*, and *c*, in the form of the latin square in Table 12.1. Then if one row were more fertile than the others, the latin square would reduce the error that this might cause. In fact, if the soil fertility was a linear function of the coordinates of the plot, the latin square arrangement would minimize the error.

Of course, the error could be reduced by subdividing the plot into a large number of pieces and planting the varieties at random. But this would make it much more difficult to sow and harvest.

Example 12.2. A smoking machine is used to test the tar content of four brands of cigarettes; the machine has four ports so that four cigarettes can be smoked simultaneously. However, the four ports might not be identical and that might affect the measurements of the tar content. Also, if four runs were made on the machine, testing one brand at a time, the humidity could change, thus affecting the results.

Show how to reduce the errors due to the different ports and the different runs by using a latin square to design the experiment.

TABLE 12.3. Design of the Smoking Experiment

		Ports			
		1 ↓	2 ↓	3 ↓	4 ↓
Runs	1 →	*c*	*b*	*a*	*d*
	2 →	*d*	*a*	*b*	*c*
	3 →	*a*	*d*	*c*	*b*
	4 →	*b*	*c*	*d*	*a*

TABLE 12.4

A	*B*	*C*	*D*	*E*
B	*A*	*E*	*C*	*D*
C	*D*	*A*	*E*	*B*
D	*E*	*B*	*A*	*C*
E	*C*	*D*	*B*	*A*

TABLE 12.5

+	..*p*	..*q*..
.		
.		
r	*A*	*B*
.		
s	*B*	*A*
.		
.		

Solution. If a, b, c, d are the four brands, we can use one of the latin squares of order 4 that we have constructed. Table 12.3 illustrates which brand should be tested at each port during each of the four runs. □

Not all latin squares can be obtained from a group table, even if we allow permutations of the rows and columns.

Example 12.3. Show that the latin square illustrated in Table 12.4 cannot be obtained from a group table.

Solution. By Corollary 4.11, all groups of order 5 are cyclic and are isomorphic to $(\mathbb{Z}_5, +)$. Suppose that the latin square in Table 12.4 could be obtained from the addition table of $\mathbb{Z}_5$. Since permutations are reversible, it follows that the rows and columns of this square could be permuted to obtain the table of $\mathbb{Z}_5$. The four elements in the left-hand top corner would be taken into four elements forming a rectangle in $\mathbb{Z}_5$, as shown in Table 12.5. Then we would have $p + r = A$, $q + r = B$, $p + s = B$, and $q + s = A$ for some $p, q, r, s \in \mathbb{Z}_5$, where $p \neq q$ and $r \neq s$. Hence $p + r = A = q + s$ and $q + r = B = p + s$. Adding, we have $p + q + 2r = p + q + 2s$ and $2r = 2s$. Therefore, $6r = 6s$, which implies that $r = s$ in $\mathbb{Z}_5$, which is a contradiction. □

ORTHOGONAL LATIN SQUARES

Suppose that in our cornfield, besides testing the yields of three varieties of corn, we also wanted to test the effects of three fertilizers on the corn. We could do

TABLE 12.6

a	b	c
c	a	b
b	c	a

A	B	C
B	C	A
C	A	B

TABLE 12.7

aA	bB	cC
cB	aC	bA
bC	cA	aB

this in the same experiment by arranging the fertilizers on the nine plots so that each of the three fertilizers was used once on each variety of corn and so that the different fertilizers themselves were arranged in a latin square of order 3.

Let a, b, c be three varieties of corn and A, B, C be three types of fertilizer. Then the two latin squares in Table 12.6 could be superimposed to form the design in Table 12.7. In this table, each variety of corn and each type of fertilizer appears exactly once in each row and in each column. Furthermore, each type of fertilizer is used exactly once with each variety of corn. This table could be used to design the experiment. For example, in the top left section of our test plot, we would plant variety a and use fertilizer A.

Two latin squares of order n are called **orthogonal** if when the squares are superimposed, each element of the first square occurs exactly once with each element of the second square. Thus the two latin squares in Table 12.6 are orthogonal.

Although it is easy to construct latin squares of any order, the construction of orthogonal latin squares can be a difficult problem. At this point the reader should try to construct two orthogonal latin squares of order 4.

Going back to our field of corn and fertilizers, could we use the same trick again to test the effect of three insecticides by choosing another latin square of order 3 orthogonal to the first two? It can be proved that it is impossible to find such a latin square (see Exercise 12.5). However, if we have four types of corn, fertilizer, and insecticide, we show using Theorem 12.5 how they could be distributed on a 4×4 plot using three latin squares of order 4 orthogonal to each other.

If $L_1, \ldots, L_r$ are latin squares of order n such that L_i is orthogonal to L_j for all $i \neq j$, then $\{L_1, \ldots, L_r\}$ is called a set of r **mutually orthogonal latin squares** of order n.

We show how to construct $n-1$ mutually orthogonal latin squares of order n from a finite field with n elements. We know that a finite field has a prime power number of elements, and we are able to construct such squares for $n = 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 17, \ldots$ etc.

Let $\mathrm{GF}(n) = \{x_0, x_1, x_2, \ldots, x_{n-1}\}$ be a finite field of order $n = p^m$, where $x_0 = 0$ and $x_1 = 1$. Let $L_1 = (a_{ij}^1)$ be the latin square of order n that is the addition table of $\mathrm{GF}(n)$. Then

$$a_{ij}^1 = x_i + x_j \quad \text{for} \quad 0 \leqslant i \leqslant n-1 \quad \text{and} \quad 0 \leqslant j \leqslant n-1.$$

Proposition 12.4. Define the squares $L_k = (a_{ij}^k)$ for $1 \leqslant k \leqslant n-1$ by

$$a_{ij}^k = x_k \cdot x_i + x_j \quad \text{for} \quad 0 \leqslant i \leqslant n-1 \quad \text{and} \quad 0 \leqslant j \leqslant n-1.$$

Then L_k is a latin square of order n for $1 \leqslant k \leqslant n-1$ based on GF(n).

Proof. The difference between two elements in the ith row is

$$\begin{aligned} a_{ij}^k - a_{iq}^k &= (x_k \cdot x_i + x_j) - (x_k \cdot x_i + x_q) \\ &= x_j - x_q \neq 0 \quad \text{if } j \neq q. \end{aligned}$$

Hence each row is a permutation of GF(n).

The difference between two elements in the jth column is

$$\begin{aligned} a_{ij}^k - a_{rj}^k &= (x_k \cdot x_i + x_j) - (x_k \cdot x_r + x_j) \\ &= x_k \cdot (x_i - x_r) \neq 0 \quad \text{if } i \neq r \quad \text{since } x_k \neq 0 \quad \text{and} \quad x_i \neq x_r. \end{aligned}$$

Hence each column is a permutation of GF(n) and L_k is a latin square of order n. □

Theorem 12.5. With the notation in Proposition 12.4, $\{L_1, L_2, \ldots, L_{n-1}\}$ is a mutually orthogonal set of latin squares of order $n = p^m$.

Proof. We have to prove that L_k is orthogonal to L_l for all $k \neq l$.

Suppose that when L_k is superimposed on L_l, the pair of elements in the (i, j)th position is the same as the pair in the (r, q)th position. That is, $(a_{ij}^k, a_{ij}^l) = (a_{rq}^k, a_{rq}^l)$ or $a_{ij}^k = a_{rq}^k$ and $a_{ij}^l = a_{rq}^l$. Hence $x_k \cdot x_i + x_j = x_k \cdot x_r + x_q$ and $x_l \cdot x_i + x_j = x_l \cdot x_r + x_q$. Subtracting, we have $(x_k - x_l) \cdot x_i = (x_k - x_l) \cdot x_r$ or $(x_k - x_l) \cdot (x_i - x_r) = 0$. Now the field GF($n$) has no zero divisors; thus either $x_k = x_l$ or $x_i = x_r$. Hence either $k = l$ or $i = r$. But $k \neq l$ and we know from Proposition 12.4 that two elements in the same row of L_k or L_l cannot be equal; therefore, $i \neq r$.

This contradiction proves that when L_k and L_l are superimposed, all the pairs of elements occurring are different. Each element of the first square appears n times and hence must occur with all the n different elements of the second square. Therefore, L_k is orthogonal to L_l, if $k \neq l$. □

If we start with $\mathbb{Z}_3$ and perform the construction above, we obtain the two mutually orthogonal latin squares of order 3 given in Table 12.8.

Example 12.6. Construct three mutually orthogonal latin squares of order 4.

TABLE 12.8. Two Orthogonal Latin Squares

L_1

0	1	2
1	2	0
2	0	1

L_2

0	1	2
2	0	1
1	2	0

TABLE 12.9. Three Mutually Orthogonal Latin Squares of Order 4

L_1

0	1	α	α^2
1	0	α^2	α
α	α^2	0	1
α^2	α	1	0

L_2

0	1	α	α^2
α	α^2	0	1
α^2	α	1	0
1	0	α^2	α

L_3

0	1	α	α^2
α^2	α	1	0
1	0	α^2	α
α	α^2	0	1

TABLE 12.10. Superimposed Latin Squares

aaa	*bbb*	*ccc*	*ddd*
bcd	*adc*	*dab*	*cba*
cdb	*dca*	*abd*	*bac*
dbc	*cad*	*bda*	*acb*

TABLE 12.11. Sixteen Court Cards

$A\spadesuit$	$K\diamondsuit$	$Q\heartsuit$	$J\clubsuit$
$Q\clubsuit$	$J\heartsuit$	$A\diamondsuit$	$K\spadesuit$
$J\diamondsuit$	$Q\spadesuit$	$K\clubsuit$	$A\heartsuit$
$K\heartsuit$	$A\clubsuit$	$J\spadesuit$	$Q\diamondsuit$

Solution. Apply the method given in Proposition 12.4 to the Galois field $\mathrm{GF}(4) = \mathbb{Z}_2(\alpha) = \{0, 1, \alpha, \alpha^2\}$, where $\alpha^2 = \alpha + 1$.

L_1 is simply the addition table for GF(4). From the way the square L_k was constructed in Proposition 12.4, we see that its rows are a permutation of the rows of L_1. Hence L_2 can be obtained by multiplying the first column of L_1 by α and then permuting the rows of L_1 so that they start with the correct element. L_3 is also obtained by permuting the rows of L_1 so that the first column is α^2 times the first column of L_1. These are illustrated in Table 12.9. □

If we write a for 0, b for 1, c for α, and d for α^2, and superimpose the three latin squares, we obtain Table 12.10. Example 12.6 also allows us to solve the parlor game of laying out the 16 cards that was mentioned at the beginning of the chapter. One solution, using the squares L_2 and L_3 in Table 12.9, is illustrated in Table 12.11.

Example 12.7. A drug company wishes to produce a new cold remedy by combining a decongestant, an antihistamine, and a pain reliever. It plans to test various combinations of three decongestants, three antihistamines, and three pain relievers on four groups of subjects each day from Monday to Thursday. Furthermore, each type of ingredient should also be compared with a placebo. Design this test so as to reduce the effects due to differences between the subject groups and the different days.

Solution. We can use the three mutually orthogonal latin squares constructed in Example 12.6 to design this experiment—see Table 12.9.

Make up the drugs given to each group using Table 12.12. The letter in the first position refers to the decongestant, the second to the antihistamine, and the third to the pain reliever. The letter a refers to a placebo, and b, c, and d refer to the three different types of ingredients. □

TABLE 12.12. Testing Three Different Drugs

		Mon.	Tues.	Wed.	Thurs.
	A	*aaa*	*bbb*	*ccc*	*ddd*
Subject	*B*	*bcd*	*adc*	*dab*	*cba*
Group	*C*	*cdb*	*dca*	*abd*	*bac*
	D	*dbc*	*cad*	*bda*	*acb*

We recognize Euler's problem of the 36 officers on parade mentioned at the beginning of the chapter as the problem of constructing two orthogonal latin squares of order 6. Euler not only conjectured that this problem was impossible to solve, but he also conjectured that it was impossible to find two orthogonal latin squares of order n, whenever $n \equiv 2 \bmod 4$.

Theorem 12.5 cannot be used to construct two such squares with order congruent to 2 modulo 4 because the only prime power of this form is 2, and then the theorem only gives one latin square. In 1899, G. Tarry, by exhaustive enumeration, proved that the problem of the 36 officers was insoluble. However, in 1959, Euler's general conjecture was shown to be false, and in fact, Bose, Shrikhande, and Parker proved that there exist at least two orthogonal latin squares of order n, for any $n > 6$. Hence Proposition 12.4 is by no means the only way of constructing orthogonal latin squares. Laywine and Mullen [40] give a comprehensive survey of all the known results on latin squares up to the time of the book's publication in 1998.

FINITE GEOMETRIES

The construction in Theorem 12.5 of $n - 1$ mutually orthogonal latin squares of order n, when n is a prime power, was first discovered by the American mathematician E. H. Moore in 1896, and was rediscovered by the Indian mathematical statistician R. C. Bose in 1938. (See Section 2.2 of Laywine and Mullin [40].) Bose also showed that there is a very close connection between orthogonal latin squares and geometries with a finite number of points and lines. These geometries are called affine planes.

An **affine plane** consists of a set, P, of **points**, together with a set, L, of subsets of P called **lines**. The points and lines must satisfy the following incidence axioms.

(i) Any two distinct points lie on exactly one line.
(ii) For each line l and point x not on l, there exists a unique line m containing x and not meeting l.
(iii) There exist three points not lying on a line.

We can define an equivalence relation of **parallelism**, $//$, on the set of lines L, by defining $l//m$ if $l = m$ or l and m contain no common point. Axiom (ii) then states that through each point there is a unique line parallel to any other line. The

points and lines in the euclidean plane $\mathbb{R}^2$ form such a geometry with an infinite number of points.

If the geometry has only a *finite* number of points, it can be shown that there exists an integer n such that the geometry contains n^2 points and $n^2 + n$ lines, and that each line contains n points, while each point lies on $n + 1$ lines. Such a finite geometry is called an **affine plane of order** n. In an affine plane of order n there are $n + 1$ parallelism classes (see Exercises 12.12 and 12.13).

Figure 12.1 shows an affine plane of order 2 in which $P = \{a, b, c, d\}$ and $L = \{\{a, b\}, \{c, d\}, \{a, c\}, \{b, c\}, \{b, d\}, \{a, d\}\}$.

Bose showed that an affine plane of order n produces a complete set of $n - 1$ mutually orthogonal latin squares of order n, and conversely, that each set of $n - 1$ mutually orthogonal latin squares of order n defines an affine plane of order n.

Theorem 12.8. There exists an affine plane of order n if and only if there exist $n - 1$ mutually orthogonal latin squares of order n.

Proof. Suppose that there exists an affine plane of order n. We coordinatize the points as follows. Take any line and label the n points as $0, 1, 2, \ldots, n - 1$. This is called the *x-axis*, and the point labeled 0 is called the *origin*. Choose any other line through the origin and label the n points $0, 1, 2, \ldots, n - 1$ with 0 at the origin. This line is called the *y-axis*. A point of the plane is said to have *coordinates* (a, b) if the unique lines through the point parallel to the y and x-axes meet the axes in points labeled a and b, respectively. This is illustrated in Figure 12.2.

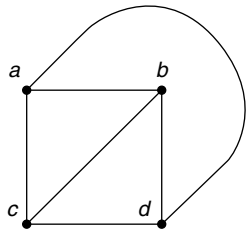


Figure 12.1. Affine plane with four points.

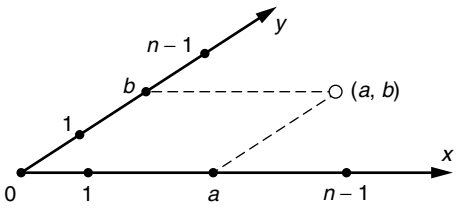


Figure 12.2. Coordinates in an affine plane.

There are n^2 ordered pairs (a, b) corresponding to the n^2 points of the plane. These points also correspond to the n^2 cells of an $n \times n$ square where (a, b) refers to the cell in the ath row and bth column. We fill these cells with numbers in $n-1$ different ways to produce $n-1$ mutually orthogonal latin squares of order n.

Consider any complete set of parallel lines that are not parallel to either axis. Label the n parallel lines $0, 1, 2, \ldots, n-1$ in any manner. Through each point, there is exactly one of these lines. In the cell (a, b) place the number of the unique line on which the point (a, b) is found. The numbers in these cells form a latin square of order n on $\{0, 1, \ldots, n-1\}$. No two numbers in the same row can be the same, because there is only one line through two points in the same row, namely, the line parallel to the x-axis. Hence each number appears exactly once in each row and, similarly, once in each column.

There are $n-1$ sets of parallelism classes that are not parallel to the axes; each of these gives rise to a latin square. These $n-1$ squares are mutually orthogonal because each line of one parallel system meets all n of the lines of any other system. Hence, when two squares are superimposed, each number of one square occurs once with each number of the second square.

Conversely, suppose that there exists a set of $n-1$ mutually orthogonal latin squares of order n. We can relabel the elements, if necessary, so that these squares are based on $S = \{0, 1, 2, \ldots, n-1\}$. We define an affine plane with S^2 as the set of points. A set of n points is said to lie on a line if there is a latin square with the same number in each of the n cells corresponding to these points, or if the n points all have one coordinate the same. It is straightforward to check that this is an affine plane of order n. □

Corollary 12.9. There exists an affine plane of order n whenever n is the power of a prime.

Proof. This follows from Theorem 12.5. □

The only known affine planes have prime power order. Because of the impossibility of solving Euler's officer problem, there are no orthogonal latin squares of order 6, and hence there is no affine plane of order 6. In 1988, by means of a massive computer search, Lam, Thiel, and Swiercz showed that there was no affine plane of order 10. See Lam [39] for the story behind this two-decade search. By Theorem 12.8, there cannot exist nine mutually orthogonal latin squares of order 10. However, two mutually orthogonal latin squares of order 10 have been found, but not three such squares. Computers have also been used to search for many sets of mutually orthogonal latin squares of low order. See Chapter 2 of Laywine and Mullen [40] for further results on mutually orthogonal latin squares.

By the method of Theorem 12.8, we can construct an affine plane of order n from the Galois field $\mathrm{GF}(n)$ whenever n is a prime power. The set of points is

$$P = \mathrm{GF}(n)^2 = \{(x, y) | x, y \in \mathrm{GF}(n)\}.$$

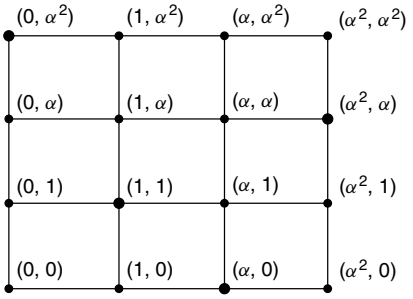


Figure 12.3. Affine plane of order 4 with the points of the line $y = \alpha x + \alpha^2$ enlarged.

It follows from Proposition 12.4 that a line consists of points satisfying a linear equation in x and y with coefficients in GF(n). The slope of a line is defined in the usual way and is an element of GF(n) or is infinite. Two lines are parallel if and only if they have the same slope.

For example, if $\mathrm{GF}(4) = \mathbb{Z}_2(\alpha) = \{0, 1, \alpha, \alpha^2\}$, the 16 points of the affine plane of order 4 are shown in Figure 12.3. The horizontal lines are of the form

$$y = \text{constant}$$

and have slope 0, whereas the vertical lines are of the form

$$x = \text{constant}$$

and have infinite slope. The line

$$y = \alpha x + \alpha^2$$

has slope α and contains the points $(0, \alpha^2)$, (1, 1), $(\alpha, 0)$ and (α^2, α). This line is parallel to the lines $y = \alpha x$, $y = \alpha x + 1$, and $y = \alpha x + \alpha$.

Given an affine plane of order n, it is possible to construct a **projective plane of order** n by adding a "line at infinity" containing $n + 1$ points corresponding to each parallelism class, so that parallel lines intersect on the line at infinity. The projective plane of order n has $n^2 + n + 1$ points and $n^2 + n + 1$ lines. Furthermore, any projective plane gives rise to an affine plane by taking one line to be the line at infinity. Hence the existence of a projective plane of order n is equivalent to the existence of an affine plane of the same order.

MAGIC SQUARES

Magic squares have been known for thousands of years, and in times when particular numbers were associated with mystical ideas, it was natural that a

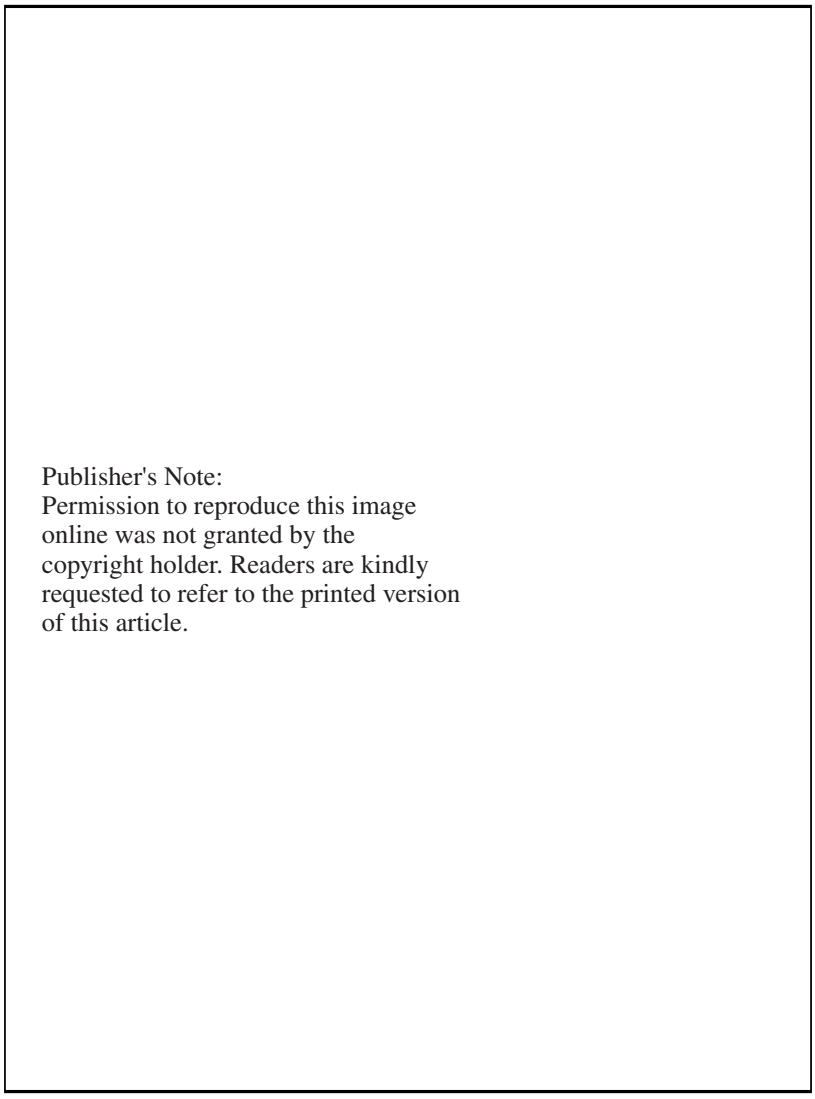


Figure 12.4. "Melancholia," an engraving by Albrecht Dürer. In the upper right there is a magic square of order 4 with the date of the engraving, 1514, in the middle of the bottom row. [Courtesy of Staatliche Museen Kupferstichkabinett, photo by Walter Steinkopf.]

square that displays such symmetry should have been deemed to have magical properties. Figure 12.4 illustrates an engraving by Dürer, made in 1514, that contains a magic square. Magic squares have no applications, and this section is included for amusement only.

A **magic square of order** n consists of the integers 1 to n^2 arranged in an $n \times n$ square array so that the row sums, column sums, and corner diagonal sums are all the same.

The sum of each row must be $n(n^2 + 1)/2$, which is $1/n$ times the sum of all the integers from 1 to n^2. For example, in Dürer's magic square of Figure 12.4, the sum of each row, column, and diagonal is 34.

It is an interesting exercise to try to construct such squares. We show how to construct some magic squares from certain pairs of orthogonal latin squares. See Ball et al. [38] and Laywine and Mullen [40] for other methods of constructing magic squares.

Let $K = (k_{ij})$ and $L = (l_{ij})$ be two orthogonal latin squares of order n on the set $S = \{0, 1, \ldots, n-1\}$. Superimpose these two squares to form a square $M = (m_{ij})$ in which the elements of M are numbers in the base n, whose first digit is taken from K and whose second digit is taken from L. That is, $m_{ij} = n \cdot k_{ij} + l_{ij}$. Since K and L are orthogonal, all possible combinations of two elements from S occur exactly once in M. In other words, all the numbers from 0 to $n^2 - 1$ occur in M.

Now add 1 to every element of M to obtain the square $M' = m'_{ij}$, where $m'_{ij} = m_{ij} + 1$.

Lemma 12.10. The square M' contains all the numbers between 1 and n^2 and is row and column magic; that is, the sums of each row and of each column are the same.

Proof. In any row or column of M, each number from 0 to $n-1$ occurs exactly once as the first digit and exactly once as the second digit. Hence the sum is

$$\begin{aligned}(0 + 1 + \cdots + n - 1)n + (0 + 1 + \cdots + n - 1)\\ = (n + 1)(n - 1)n/2 = n(n^2 - 1)/2.\end{aligned}$$

Therefore, each row or column sum of M' is $n(n^2 - 1)/2 + n = n(n^2 + 1)/2$.□

Example 12.11. Construct the square M' from the two orthogonal latin squares, K and L, in Table 12.13.

Solution. Table 12.13 illustrates the superimposed square M in base 3 and in base 10. By adding one to each element, we obtain the magic square M'. □

Theorem 12.12. If K and L are orthogonal latin squares of order n on the set $\{0, 1, 2, \ldots, n-1\}$ and the sum of each of the diagonals of K and L is $n(n-1)/2$, then the square M' derived from K and L is a magic square of order n.

TABLE 12.13. Construction of a Magic Square of Order 3

1	2	0
0	1	2
2	0	1

K

0	2	1
2	1	0
1	0	2

L

10	22	01
02	11	20
21	00	12

M (in base 3)

3	8	1
2	4	6
7	0	5

M (in base 10)

4	9	2
3	5	7
8	1	6

M'

Proof. Lemma 12.10 shows that the sum of each row and each column is $n(n^2+1)/2$. A similar argument shows that the sum of each diagonal is also $n(n^2+1)/2$. □

There are two common ways in which the sum of the diagonal elements of K and L can equal $n(n-1)/2$.

(i) The diagonal is a permutation of $\{0, 1, \ldots, n-1\}$.

(ii) If n is odd, every diagonal element is $(n-1)/2$.

Both these situations occur in the squares K and L of Table 12.13; thus the square M', which is constructed from these, is a magic square.

Example 12.13. Construct a magic square of order 4 from two orthogonal latin squares in Table 12.9.

Solution. By replacing $0, 1, \alpha, \alpha^2$ by 0, 1, 2, 3, in any order, the squares L_2 and L_3 in Table 12.9 satisfy the conditions of Theorem 12.12, because the

TABLE 12.14. Construction of a Magic Square of Order 4

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

L_2'

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

L_3'

03	12	20	31
21	30	02	13
32	23	11	00
10	01	33	22

M (in base 4)

4	7	9	14
10	13	3	8
15	12	6	1
5	2	16	11

M'

diagonal elements are all different. However, L_1 will not satisfy the conditions of Theorem 12.12, whatever substitutions we make. In L_2, replace $0, 1, \alpha, \alpha^2$ by 0, 1, 2, 3, respectively, and in L_3 replace $0, 1, \alpha, \alpha^2$ by 3, 2, 0, 1, respectively, to obtain the squares L_2' and L_3' in Table 12.14. Combine these to obtain the square M with entries in base 4. Add 1 to each entry and convert to base 10 to obtain the magic square M' in Table 12.14. □

EXERCISES

12.1. Construct a latin square of order 7 on $\{a, b, c, d, e, f, g\}$.

12.2. Construct four mutually orthogonal latin squares of order 5.

12.3. Construct four mutually orthogonal latin squares of order 8.

12.4. Construct two mutually orthogonal latin squares of order 9.

12.5. Prove that there are at most $(n - 1)$ mutually orthogonal latin squares of order n. (You can always relabel each square so that the first rows are the same.)

12.6. Let $L = (l_{ij})$ be a latin square of order l on $\{1, 2, \ldots, l\}$ and $M = (m_{ij})$ be a latin square of order m on $\{1, 2, \ldots, m\}$. Describe how to construct a latin square of order lm on $\{1, 2, \ldots, l\} \times \{1, 2, \ldots, m\}$ from L and M.

12.7. Is the latin square of Table 12.15 the multiplication table for a group of order 6 with identity A?

TABLE 12.15

A	B	C	D	E	F
B	A	F	E	C	D
C	F	B	A	D	E
D	C	E	B	F	A
E	D	A	F	B	C
F	E	D	C	A	B

12.8. A chemical company wants to test a chemical reaction using seven different levels of catalyst, a, b, c, d, e, f, g. In the manufacturing process, the raw material comes from the previous stage in batches, and the catalyst must be added immediately. If there are seven reactors, A, B, C, D, E, F, G, in which the catalytic reaction can take place, show how to design the experiment using seven batches of raw material so as to minimize the effect of the different batches and of the different reactors.

12.9. A supermarket wishes to test the effect of putting cereal on four shelves at different heights. Show how to design such an experiment lasting four weeks and using four brands of cereal.

12.10. A manufacturer has five types of toothpaste. He would like to test these on five subjects by giving each subject a different type each week for five weeks. Each type of toothpaste is identified by a different color—red, blue, green, white, or purple—and the manufacturer changes the color code each week to reduce the psychological effect of the color. Show how to design this experiment.

12.11. Quality control would like to find the best type of music to play to its assembly line workers in order to reduce the number of faulty products. As an experiment, a different type of music is played on four days in a week, and on the fifth day no music at all is played. Design such an experiment to last five weeks that will reduce the effect of the different days of the week.

12.12. The relation of parallelism, //, on the set of lines of an affine plane is defined by $l//m$ if and only if $l = m$ or $l \cap m = \emptyset$. Prove that // is an equivalence relation.

12.13. Let P be the set of points and L be the set of lines of a finite affine plane.
- **(a)** Show that the number of points on a line l equals the number of lines in any parallelism class not containing l.
- **(b)** Deduce that all the lines contain the same number of points.
- **(c)** If each line contains n points, show that the plane contains n^2 points and $n^2 + n$ lines, each point lying on $n + 1$ lines. Show also that there are $n + 1$ parallelism classes.

12.14. Find all the lines in the affine plane of order 3 whose point set is $\mathbb{Z}_3^2$.

Exercises 12.15 to 12.17 refer to the affine plane of order 9 obtained from $GF(9) = \mathbb{Z}_3(\alpha)$, *where* $\alpha^2 + 1 = 0$.

12.15. Find the line through $(2\alpha, 1)$ that is parallel to the line $y = \alpha x + 2 + \alpha$.

12.16. Find the point of intersection of the lines $y = x + \alpha$ and $y = (\alpha + 1)x + 2\alpha$.

12.17. Find the equation of the line through $(0, 2\alpha)$ and $(2, \alpha + 1)$.

12.18. Prove that a magic square of order 3 must have 5 at its center.

12.19. Prove that 1 cannot be a corner element of a magic square of order 3.

12.20. How many different magic squares of order 3 are there?

12.21. How many essentially different magic squares of order 3 are there, that is, magic squares that cannot be obtained from each other by a symmetry of the square?

12.22. Is there a magic square of order 2?

12.23. Find two magic squares of order 4 different from the square in Example 12.18.

12.24. Find a magic square of order 5.

12.25. Find a magic square of order 8.

12.26. Can you construct a magic square with the present year in the last two squares of the bottom row?

13

GEOMETRICAL CONSTRUCTIONS

The only geometric instruments used by the ancient Greeks were a straightedge and a compass. They did not possess reliable graduated rulers or protractors. However, with these two instruments, they could still perform a wide variety of constructions; they could divide a line into any number of equal parts, and they could bisect angles and construct parallel and perpendicular lines. There were three famous problems that the Greeks could not solve using these methods: (1) *duplication of the cube*; that is, given one edge of a cube, construct the edge of a cube whose volume is double that of the given cube; (2) *trisection of any given angle*; and (3) *squaring of the circle*; that is, given any circle, construct a square whose area is the same as that of the circle. For centuries, the solution to these problems eluded mathematicians, despite the fact that large prizes were offered for their discovery.

It was not until the nineteenth century that mathematicians suspected and, in fact, proved that these constructions were impossible. In the beginning of that century, nonexistence proofs began appearing in algebra; it was proved that the general polynomial equation of degree 5 could not be solved in terms of nth roots and rational operations. Similar algebraic methods were then applied to these geometric problems. The geometric problems could be converted into algebraic problems by determining which multiples of a given length could be constructed using only straightedge and compass. Some of the classical constructions are illustrated in Figure 13.1.

CONSTRUCTIBLE NUMBERS

We are interested in those lengths that can be constructed from a given length. For convenience, we choose our unit of length to be the given length. We see

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

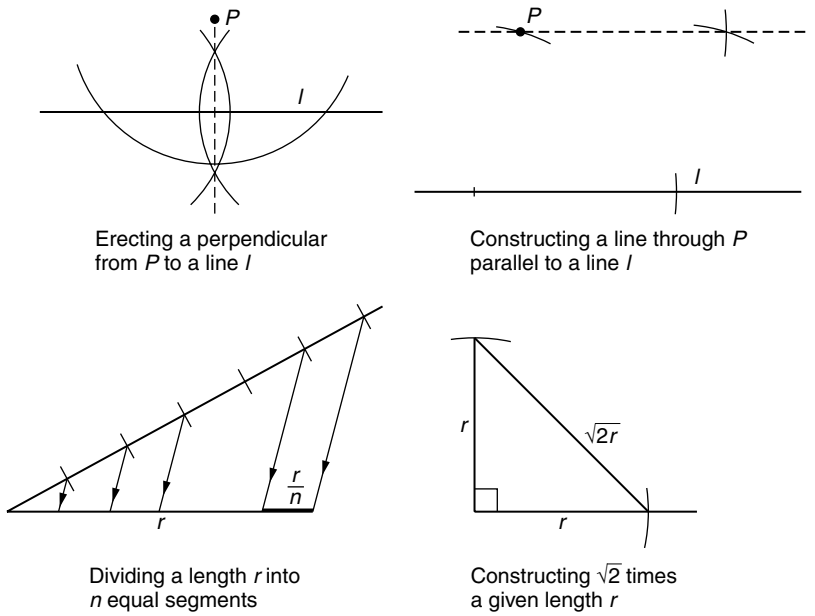


Figure 13.1. Geometrical constructions using straightedge and compass.

that we can divide a length into any number of equal parts, and hence we can construct any rational multiple. However, we can do more than this; we can construct irrational multiples such as $\sqrt{2}$ by using right-angled triangles.

Given any line segment in the plane, choose rectangular coordinates so that the line's end points are (0, 0) and (1, 0). Any point in the plane that can be constructed from this line segment by using a straightedge and compass is called a **constructible point**. A real number is called **constructible** if it occurs as one coordinate of a constructible point. Points can be constructed by performing the following allowable operations a finite number of times. We can:

1. Draw a line through two previously constructed points.
2. Draw a circle with center at a previously constructed point and with radius equal to the distance between two previously constructed points.
3. Mark the point of intersection of two straight lines.
4. Mark the points of intersection of a straight line and a circle.
5. Mark the points of intersection of two circles.

Theorem 13.1. The set of constructible numbers, K, is a subfield of $\mathbb{R}$.

Proof. K is a subset of $\mathbb{R}$, so we have to show that it is a field. That is, if $a, b \in K$, we have to show that $a + b, a - b, ab$, and if $b \neq 0, a/b \in K$.

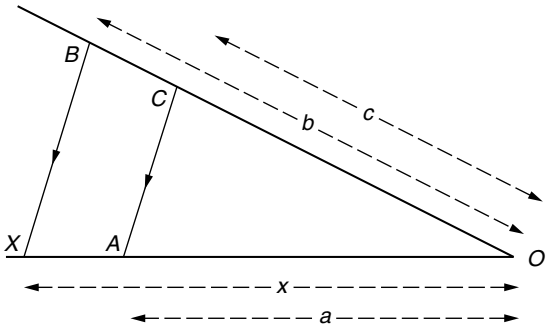


Figure 13.2. Constructing products and quotients.

If $a, b \in K$, we can mark off lengths a and b on a line to construct lengths $a + b$ and $a - b$.

If $a, b, c \in K$, mark off a segment OA of length a on one line and mark off segments OB and OC of length b and c on another line through O as shown in Figure 13.2. Draw a line through B parallel to CA and let it meet OA in X. Triangles OAC and OXB are similar, and if $OX = x$, then $x/a = b/c$ and so $x = ab/c$.

By taking $c = 1$, we can construct ab, and by taking $b = 1$, we can construct a/c. Hence K is a subfield of $\mathbb{R}$. □

Corollary 13.2. K is an extension field of $\mathbb{Q}$.

Proof. Since $1 \in K$ and sums and differences of constructible numbers are constructible, it follows that $\mathbb{Z} \subseteq K$. Since quotients of constructible numbers are constructible, $\mathbb{Q} \subseteq K$. □

Proposition 13.3. If $k \in K$ and $k > 0$, then $\sqrt{k} \in K$.

Proof. Mark off segments AB and BC of lengths k and 1 on a line. Draw the circle with diameter AC and construct the perpendicular to AC at B as shown in Figure 13.3. Let it meet the circle at D and E. Then, by a standard theorem in geometry, $AB \cdot BC = DB \cdot BE$: thus $BD = BE = \sqrt{k}$. □

Example 13.4. $\sqrt[4]{2}$ is constructible.

Solution. We apply the construction of Proposition 13.3 twice to construct $\sqrt{2}$ and then $\sqrt{\sqrt{2}} = \sqrt[4]{2}$. □

We can construct any number that can be written in terms of rational numbers, $+, -, \cdot, \div$, and $\sqrt{\ }$ signs. For example, the numbers $1 + 4\sqrt{5}$, $\sqrt{2} + \sqrt{4/5}$, and $\sqrt{3 - \sqrt{7}}$ are all constructible. If k_1 is a positive rational number, all the elements

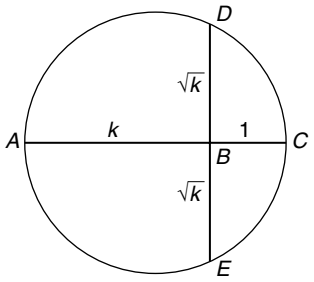


Figure 13.3. Constructing square roots.

of the extension field $K_1 = \mathbb{Q}(\sqrt{k_1})$ are constructible. K_1 has degree 1 or 2 over $\mathbb{Q}$ depending on whether $\sqrt{k_1}$ is rational or irrational. If k_2 is a positive element of K_1, all the elements of $K_2 = K_1(\sqrt{k_2}) = \mathbb{Q}(\sqrt{k_1}, \sqrt{k_2})$ are constructible, and $[K_2 : K_1] = 1$ or 2, depending on whether or not $\sqrt{k_2}$ is an element of K_1. We now show that every constructible number lies in a field obtained by repeating the extensions above.

Theorem 13.5. The number α is constructible if and only if there exists a sequence of real fields $K_0, K_1, \ldots, K_n$ such that $\alpha \in K_n \supseteq K_{n-1} \supseteq \cdots \supseteq K_0 = \mathbb{Q}$ and $[K_i : K_{i-1}] = 2$ for $1 \leqslant i \leqslant n$.

Proof. Suppose that $\alpha \in K_n \supseteq K_{n-1} \supseteq \cdots \supseteq K_0 = \mathbb{Q}$, where $[K_i : K_{i-1}] = 2$. By Proposition 11.16, $K_i = K_{i-1}(\sqrt{\gamma_{i-1}})$ for $\gamma_{i-1} \in K_{i-1}$, and since K_i is real, $\gamma_{i-1} > 0$. Therefore, by repeated application of Proposition 13.3, it can be shown that every element of K_n is constructible.

Conversely, suppose that $\alpha \in K$; thus α appears as the coordinate of a point constructible from (0, 0) and (1, 0) by a finite number of the operations 1 to 5 preceding Theorem 13.1. We prove the result by induction on m, the number of constructible numbers used in reaching α.

Suppose that $X_k = \{x_1, \ldots, x_k\}$ is a set of numbers that have already been constructed, that is, have appeared as coordinates of constructible points. When the next number x_{k+1} is constructed, we show that $[\mathbb{Q}(X_{k+1}) : \mathbb{Q}(X_k)] = 1$ or 2, where $\mathbb{Q}(X_{k+1}) = \mathbb{Q}(x_1, \ldots, x_k, x_{k+1})$.

We first show that if we perform either operation 1 or 2 using previously constructed numbers in X_k, the coefficients in the equation of the line or circle remain in $\mathbb{Q}(X_k)$. If we perform operation 3, the newly constructed numbers remain in $\mathbb{Q}(X_k)$, and if we perform operation 4 or 5, the newly constructed numbers are either in $\mathbb{Q}(X_k)$ or an extension field of degree 2 over $\mathbb{Q}(X_k)$.

Operation 1. The line through (α_1, β_1) and (α_2, β_2) is $(y - \beta_1)/(\beta_2 - \beta_1) = (x - \alpha_1)/(\alpha_2 - \alpha_1)$, and if $\alpha_1, \alpha_2, \beta_1, \beta_2 \in X_k$, the coefficients in the equation of this line lie in $\mathbb{Q}(X_k)$.

Operation 2. The circle with center (α_1, β_1) and radius equal to the distance from (α_2, β_2) to (α_3, β_3) is $(x-\alpha_1)^2 + (y-\beta_1)^2 = (\alpha_2-\alpha_3)^2 + (\beta_2-\beta_3)^2$, and all the coefficients in this equation lie in $\mathbb{Q}(X_k)$.

Operation 3. Let $\alpha_{ij}, \beta_j \in \mathbb{Q}(X_k)$. Then the lines

$$\alpha_{11}x + \alpha_{12}y = \beta_1$$
$$\alpha_{21}x + \alpha_{22}y = \beta_2$$

meet in the point (x, y), where using Cramer's rule,

$$x = \det\begin{pmatrix} \beta_1 & \alpha_{12} \\ \beta_2 & \alpha_{22} \end{pmatrix} \Big/ \det\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$$

and

$$y = \det\begin{pmatrix} \alpha_{11} & \beta_1 \\ \alpha_{21} & \beta_2 \end{pmatrix} \Big/ \det\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$$

as long as they are not parallel. Both of these coordinates are in $\mathbb{Q}(X_k)$.

Operation 4. To obtain the points of intersection of a circle and line with coefficients in $\mathbb{Q}(X_k)$, we eliminate y from the equations to obtain an equation of the form

$$\alpha x^2 + \beta x + \gamma = 0 \quad \text{where} \quad \alpha, \beta, \gamma \in \mathbb{Q}(X_k).$$

The line and circle intersect if $\beta^2 - 4\alpha\gamma \geqslant 0$ and the x coordinates of the intersection points are $x = \dfrac{-\beta \pm \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}$, which are in $\mathbb{Q}(X_k)$ or $\mathbb{Q}(X_k)(\sqrt{\beta^2 - 4\alpha\gamma})$. Similarly, the y coordinates are in $\mathbb{Q}(X_k)$ or in an extension field of degree 2 over $\mathbb{Q}(X_k)$.

Operation 5. The intersection of the two circles

$$x^2 + y^2 + \alpha_1 x + \beta_1 y + \gamma_1 = 0$$
$$x^2 + y^2 + \alpha_2 x + \beta_2 y + \gamma_2 = 0$$

is the same as the intersection of one of them with the line

$$(\alpha_1 - \alpha_2)x + (\beta_1 - \beta_2)y + (\gamma_1 - \gamma_2) = 0.$$

This is now the same situation as in operation (4).

Initially, $m = 2$, $X_2 = \{0, 1\}$, and $\mathbb{Q}(X_2) = \mathbb{Q}$. It follows by induction on m, the number of constructible points used, that

$$\alpha \in \mathbb{Q}(X_m) \supseteq \mathbb{Q}(X_{m-1}) \supseteq \cdots \supseteq \mathbb{Q}(X_3) \supseteq \mathbb{Q}(X_2) = \mathbb{Q},$$

where $[\mathbb{Q}(X_{k+1}) : \mathbb{Q}(X_k)] = 1$ or 2 for $2 \leqslant k \leqslant m-1$. Furthermore, each extension field $\mathbb{Q}(X_k)$ is a subfield of $\mathbb{R}$ because $\mathbb{Q}$ and X_k are sets of real numbers. By dropping each field $\mathbb{Q}(X_i)$ that is a trivial extension of $\mathbb{Q}(X_{i-1})$, it follows that

$$\alpha \in K_n \supseteq K_{n-1} \supseteq \cdots \supseteq K_0 = \mathbb{Q}$$

where $[K_i : K_{i-1}] = 2$ for $1 \leqslant i \leqslant n$. □

Corollary 13.6. If α is constructible, then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^r$ for some $r \geqslant 0$.

Proof. If α is constructible, then $\alpha \in K_n \supseteq K_{n-1} \supseteq \cdots \supseteq K_0 = \mathbb{Q}$, where K_i is an extension field of degree 2 over K_{i-1}. By Theorem 11.6,

$$\begin{aligned}[K_n : \mathbb{Q}(\alpha)][\mathbb{Q}(\alpha) : \mathbb{Q}] &= [K_n : \mathbb{Q}] \\ &= [K_n : K_{n-1}][K_{n-1} : K_{n-2}] \cdots [K_1 : \mathbb{Q}] = 2^n.\end{aligned}$$

Hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] | 2^n$; thus $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^r$ for some $r \geqslant 0$. □

Corollary 13.7. If $[\mathbb{Q}(\alpha) : \mathbb{Q}] \neq 2^r$ for some $r \geqslant 0$, then α is not constructible.

Corollary 13.6 does not give a sufficient condition for α to be constructible, as shown in Example 13.17 below.

Example 13.8. Can a root of the polynomial $x^5 + 4x + 2$ be constructed using straightedge and compass?

Solution. Let α be a root of $x^5 + 4x + 2$. By Eisenstein's criterion, $x^5 + 4x + 2$ is irreducible over $\mathbb{Q}$; thus, by Corollary 11.12, $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 5$. Since 5 is not a power of 2, it follows from Corollary 13.7 that α is not constructible. □

Example 13.9. Can a root of the polynomial $x^4 - 3x^2 + 1$ be constructed using straightedge and compass?

Solution. Solving the equation $x^4 - 3x^2 + 1 = 0$, we obtain $x^2 = (3 \pm \sqrt{5})/2$ and $x = \pm\sqrt{(3 \pm \sqrt{5})/2}$. It follows from Theorem 13.5 that all these roots can be constructed. □

DUPLICATING A CUBE

Let l be the length of the sides of a given cube so that its volume is l^3. A cube with double the volume will have sides of length $\sqrt[3]{2}\, l$.

Proposition 13.10. $\sqrt[3]{2}$ is not constructible.

Proof. $\sqrt[3]{2}$ is a root of $x^3 - 2$ which, by the rational roots theorem (Theorem 9.25), is irreducible over $\mathbb{Q}$. Hence, by Corollary 11.12, $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$ so, by Corollary 13.7, $\sqrt[3]{2}$ is not constructible. □

Since we cannot construct a length of $\sqrt[3]{2}\, l$ starting with a length l, the ancient problem of duplicating the cube is insoluble.

TRISECTING AN ANGLE

Certain angles can be trisected using straightedge and compass. For example, $\pi, \pi/2, 3\pi/4$ can be trisected because $\pi/3, \pi/6$, and $\pi/4$ can be constructed. However, we show that not all angles are trisectable by proving that $\pi/3$ cannot be trisected.

If we are given the angle ϕ, we can drop a perpendicular from a point a unit distance from the angle to construct the lengths $\cos\phi$ and $\sin\phi$, as shown in Figure 13.4. Conversely, if either $\cos\phi$ or $\sin\phi$ is constructible, it is possible to construct the angle ϕ. Hence, if we are given an angle ϕ, we can construct all numbers in the extension field $\mathbb{Q}(\cos\phi)$. Of course, if $\cos\phi \in \mathbb{Q}$, then $\mathbb{Q}(\cos\phi) = \mathbb{Q}$.

We can now consider those numbers that are constructible from $\mathbb{Q}(\cos\phi)$. This notion of constructibility is similar to our previous notion, and similar results hold, except that the starting field is $\mathbb{Q}(\cos\phi)$ instead of $\mathbb{Q}$.

Theorem 13.11. The angle ϕ can be trisected if and only if the polynomial $4x^3 - 3x - \cos\phi$ is reducible over $\mathbb{Q}(\cos\phi)$.

Proof. Let $\theta = \phi/3$. The angle θ can be constructed from ϕ if and only if $\cos\theta$ can be constructed from $\cos\phi$. It follows from De Moivre's theorem and the binomial theorem that

$$\cos\phi = \cos 3\theta = 4\cos^3\theta - 3\cos\theta.$$

Hence $\cos\theta$ is a root of $f(x) = 4x^3 - 3x - \cos\phi$.

If $f(x)$ is reducible over $\mathbb{Q}(\cos\phi)$, then $\cos\theta$ is a root of a polynomial of degree 1 or 2 over $\mathbb{Q}(\cos\phi)$; thus $[\mathbb{Q}(\cos\phi, \cos\theta) : \mathbb{Q}(\cos\phi)] = 1$ or 2. Hence, by Propositions 11.16 and 13.3, $\cos\theta$ is constructible from $\mathbb{Q}(\cos\phi)$.

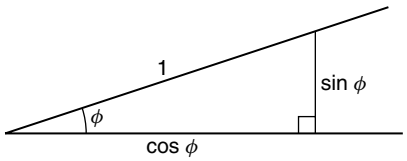


Figure 13.4. Constructing $\sin\phi$ and $\cos\phi$ from the angle ϕ.

If $f(x)$ is irreducible over $\mathbb{Q}(\cos\phi)$, then $[\mathbb{Q}(\cos\phi, \cos\theta) : \mathbb{Q}(\cos\phi)] = 3$, and it follows, by a proof similar to that of Theorem 13.5, that $\cos\theta$ cannot be constructed from $\mathbb{Q}(\cos\phi)$ by using straightedge and compass. □

Corollary 13.12. If $\cos\phi \in \mathbb{Q}$, then the angle ϕ can be trisected if and only if $4x^3 - 3x - \cos\phi$ is reducible over $\mathbb{Q}$.

For example, if $\phi = \pi/2$, then ϕ can be trisected because the polynomial $4x^3 - 3x + 0$ is reducible over $\mathbb{Q}$.

Proposition 13.13. $\pi/3$ cannot be trisected by straightedge and compass.

Proof. The polynomial $f(x) = 4x^3 - 3x - \cos(\pi/3) = 4x^3 - 3x - \frac{1}{2}$. Now, by the rational roots theorem (Theorem 9.25), the only possible roots of $2f(x) = 8x^3 - 6x - 1$ are $\pm 1, \pm\frac{1}{2}, \pm\frac{1}{4}$, or $\pm\frac{1}{8}$. We see from the graph of $f(x)$ in Figure 13.5 that none of these are roots, except possibly $-\frac{1}{4}$ or $-\frac{1}{8}$. However, $f(-\frac{1}{4}) = \frac{3}{16}$ and $f(-\frac{1}{8}) = -\frac{17}{128}$; thus $f(x)$ has no rational roots. Hence $f(x)$ is irreducible over $\mathbb{Q}$, and by Corollary 13.12, $\pi/3$ cannot be trisected by straightedge and compass. □

Example 13.14. Archimedes showed that, if we are allowed to mark our straightedge, it *is* possible to trisect any angle.

Construction. Let AOB be the angle ϕ we are to trisect. Draw a circle with center O and any radius r and let this circle meet OA and OB in P and Q. Mark two points X and Y on our straightedge of distance r apart. Now move this straightedge through Q, keeping X on OA until Y lies on the circle, as shown in Figure 13.6. Then we claim that the angle OXY is $\phi/3$, and hence the angle AOB is trisected.

Solution. Let angle $OXY = \theta$. Since triangle XYO is isosceles, the angle $XOY = \theta$. Now

$$\text{angle } OYQ = \text{angle } OXY + \text{angle } XOY = 2\theta.$$

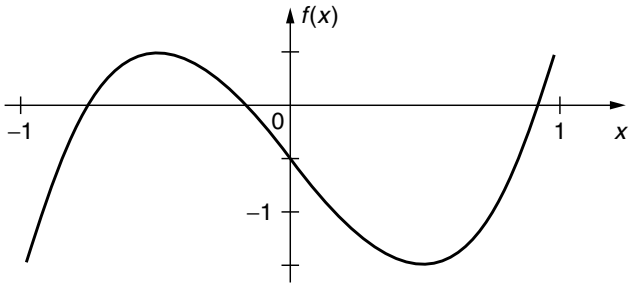


Figure 13.5. Graph of $f(x) = 4x^3 - 3x - \frac{1}{2}$.

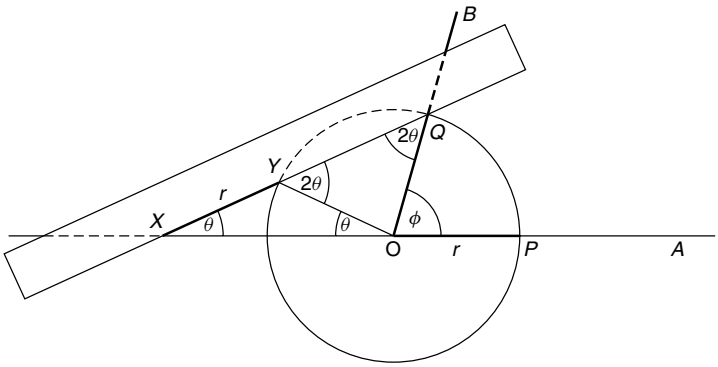


Figure 13.6. Trisection of the angle ϕ using a marked ruler.

Triangle YOQ is isosceles, so angle $OQY = 2\theta$. Also,

$$\phi = \text{angle } AOB = \text{angle } OXQ + \text{angle } OQX = \theta + 2\theta = 3\theta.$$

Hence $\theta = \phi/3$. □

SQUARING THE CIRCLE

Given any circle of radius r, its area is πr^2, so that a square with the same area has sides of length $\sqrt{\pi}\, r$. We can square the circle if and only if $\sqrt{\pi}$ is constructible.

Proposition 13.15. $[\mathbb{Q}(\sqrt{\pi}) : \mathbb{Q}]$ is infinite, and hence $\sqrt{\pi}$ is not constructible.

Proof. The proof of this depends on the fact that π is transcendental over $\mathbb{Q}$; that is, π does not satisfy any polynomial equation with rational coefficients. This was mentioned in Chapter 11, and a proof is given in Stewart [35].

$\mathbb{Q}(\pi)$ is a subfield of $\mathbb{Q}(\sqrt{\pi})$ because $\pi = (\sqrt{\pi})^2 \in \mathbb{Q}(\sqrt{\pi})$. Since π is transcendental, $\pi, \pi^2, \pi^3, \ldots$ are linearly independent over $\mathbb{Q}$, and $[\mathbb{Q}(\pi) : \mathbb{Q}]$ is infinite. Therefore,

$$[\mathbb{Q}(\sqrt{\pi}) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{\pi}) : \mathbb{Q}(\pi)][\mathbb{Q}(\pi) : \mathbb{Q}]$$

is also infinite. Hence, by Corollary 13.7, $\sqrt{\pi}$ is not constructible. □

Hence the circle cannot be squared by straightedge and compass.

CONSTRUCTING REGULAR POLYGONS

Another problem that has been of great interest to mathematicians from the time of the ancient Greeks is that of constructing a regular n-gon, that is, a regular polygon with n sides. This is equivalent to constructing the angle $2\pi/n$ or the

number $\cos(2\pi/n)$. The Greeks knew how to construct regular polygons with three, four, five, and six sides, but were unable to construct a regular 7-gon.

It is well known how to construct an equilateral triangle and a square using straightedge and compass. We proved in Example 11.15 that $\cos(2\pi/5) = (\sqrt{5}-1)/4$; thus a regular pentagon can be constructed. Furthermore, if a regular n-gon is constructible, so is a regular $2n$-gon, because angles can be bisected using straightedge and compass. Proposition 13.13 shows that $\pi/9$ cannot be constructed; hence $2\pi/9$ and a regular 9-gon cannot be constructed.

In 1796, at the age of 19, Gauss discovered that a regular 17-gon could be constructed and later showed the only regular n-gons that are constructible are the ones for which $n = 2^k p_1 \cdots p_r$, where $k \geqslant 0$ and $p_1, \ldots, p_r$ are distinct primes of the form $2^{2^m}+1$. Prime numbers of the form $2^{2^m}+1$ are called **Fermat primes**. Pierre de Fermat (1601–1665) conjectured that all numbers of the form $2^{2^m}+1$ are prime. When $m = 0, 1, 2, 3$, and 4, the numbers are 3, 5, 17, 257, and 65,537, respectively, and they are all prime. However, in 1732, Euler discovered that $2^{2^5}+1$ is divisible by 641. Computers have checked many of these numbers for $m > 5$, and they have all been composite. In fact, no more Fermat primes are known today.

A complete proof of Gauss's result is beyond the scope of this book, since it requires more group and field theory than we have covered (see Stewart [35]). However, we can prove the following.

Theorem 13.16. If p is a prime for which a regular p-gon is constructible, then p is a Fermat prime.

Proof. Let $\xi_p = \cos(2\pi/p) + i\sin(2\pi/p)$, a pth root of unity. If a regular p-gon can be constructed, $\cos(2\pi/p)$ and $\sin(2\pi/p)$ are constructible numbers and $[\mathbb{Q}(\cos(2\pi/p), \sin(2\pi/p)) : \mathbb{Q}] = 2^r$ for some integer r. Hence

$$[\mathbb{Q}(\cos(2\pi/p), \sin(2\pi/p), i) : \mathbb{Q}] = 2^{r+1}.$$

Now $\mathbb{Q}(\xi_p) \subseteq \mathbb{Q}(\cos(2\pi/p), \sin(2\pi/p), i)$ and so, by Theorem 11.6, $[\mathbb{Q}(\xi_p) : \mathbb{Q}] = 2^k$ for some integer $k \leqslant r+1$.

The pth root of unity, ξ_p, is a root of the cyclotomic polynomial $\phi(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$ which, by Example 9.31, is irreducible over $\mathbb{Q}$. Hence $[\mathbb{Q}(\xi_p) : \mathbb{Q}] = p-1$, and therefore $p - 1 = 2^k$.

The number $p = 2^k + 1$ is a prime only if $k = 0$ or k is a power of 2. Suppose that k contains an odd factor $b > 1$ and that $k = a \cdot b$. Then $2^a + 1$ divides $(2^a)^b + 1$, since $x + 1$ divides $x^b + 1$ if b is odd. Hence $2^{ab} + 1$ cannot be prime.

The case $p = 2$ gives rise to the degenerate 2-gon. Otherwise, p is a Fermat prime, $2^{2^m}+1$, for some integer $m \geqslant 0$. □

NONCONSTRUCTIBLE NUMBER OF DEGREE 4

This next example shows that Corollary 13.6 does not give a sufficient condition for a number to be constructible.

Example 13.17. There is a real root r_i, of the irreducible polynomial $x^4 - 4x + 2$, that is not constructible, even though $[\mathbb{Q}(r_i) : \mathbb{Q}] = 2^2$.

Solution. By Eisenstein's criterion, $x^4 - 4x + 2$ is irreducible over $\mathbb{Q}$, so that $[\mathbb{Q}(r_i) : \mathbb{Q}] = 4$ for each root r_i. However, we can factor this polynomial into two quadratics over $\mathbb{R}$, say

$$x^4 - 4x + 2 = (x^2 + ax + b)(x^2 + cx + d).$$

Comparing coefficients, and then using equation (i), we have

$$\begin{array}{lllll} \text{(i)} & 0 = a + c & \text{and} & c = -a \\ \text{(ii)} & 0 = b + d + ac & \text{and} & b + d = a^2 \\ \text{(iii)} & -4 = bc + ad & \text{and} & -4 = a(d - b) \\ \text{(iv)} & 2 = bd. \end{array}$$

Let $t = b + d$, so that $16 = a^2\{(b + d)^2 - 4bd\} = t(t^2 - 8)$. This number t satisfies the equation

(v) $t^3 - 8t - 16 = 0$.

By Theorem 9.25, this equation (v) has no rational roots; thus $t^3 - 8t - 16$ is irreducible over $\mathbb{Q}$. We see from Figure 13.7 that the equation does have a real root ρ between 3 and 4, and the coefficients a, b, c, d can be expressed in terms of ρ.

Either a or c is positive. Without loss of generality suppose that $c > 0$; thus $b + d = t = \rho$, $a = -c = -\sqrt{\rho}$, and $d - b = 4/\sqrt{\rho}$. Therefore, we have $b = \rho/2 - 2/\sqrt{\rho}$ and $d = \rho/2 + 2/\sqrt{\rho}$, and the roots of $x^2 + ax + b$ are

$$\frac{-a \pm \sqrt{a^2 - 4b}}{2} = \frac{1}{2}\left[\sqrt{\rho} \pm \sqrt{-\rho + \frac{8}{\sqrt{\rho}}}\right],$$

which are real, since $\rho < 4$. These are the roots r_1 and r_2 in Figure 13.8.

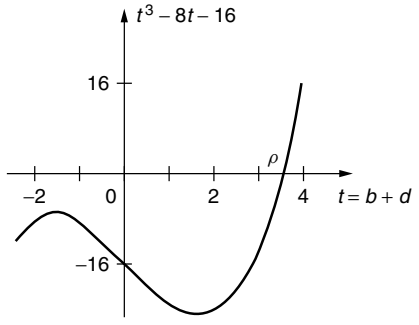


Figure 13.7. Graph of $t^3 - 8t - 16$.

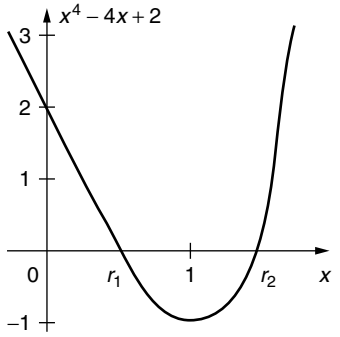


Figure 13.8. Graph of $x^4 - 4x + 2$.

If both these roots of $x^4 - 4x + 2$ are constructible, then $(r_1 + r_2)^2 = \rho$ is also constructible. But this is impossible, since ρ is a root of the irreducible polynomial $t^3 - 8t - 16$ and $[\mathbb{Q}(\rho) : \mathbb{Q}] = 3$.

Hence $x^4 - 4x + 2$ has a real root that is not constructible. □

This example was adapted from the article by Kalmanson [42].

EXERCISES

For Exercises 13.1 to 13.6, which of the numbers are constructible?

13.1. $\sqrt[4]{5 + \sqrt{2}}$.

13.2. $\sqrt[6]{2}$.

13.3. $\dfrac{2}{1 + \sqrt{7}}$.

13.4. $(1 - 4\sqrt{7})^3$.

13.5. $1 - \sqrt[5]{27}$.

13.6. $\sqrt[3]{7 - 5\sqrt{2}}$.

13.7. Is $\mathbb{Q}(\cos\phi) = \mathbb{Q}(\sin\phi)$ for every angle ϕ?

13.8. If $\tan\phi$ is constructible, show how to construct the angle ϕ.

13.9. Prove that all the constructible numbers are algebraic over $\mathbb{Q}$.

For the values of $\cos\phi$ *given in Exercises 13.10 to 13.13, determine whether you can trisect the angle* ϕ *by straightedge and compass.*

13.10. $\cos\phi = 1/4$.

13.11. $\cos\phi = -9/16$.

13.12. $\cos\phi = 1/\sqrt{2}$.

13.13. $\cos\phi = \sqrt{2}/8$.

13.14. By writing $\pi/15$ in terms of $\pi/5$ and $\pi/3$, show that it is possible to trisect $\pi/5$ and also possible to construct a regular 15-gon.

13.15. Can $\pi/7$ be trisected?

13.16. Construct a regular pentagon using straightedge and compass only.

13.17. Prove that $\cos(2\pi/7)$ is a root of $8x^3 + 4x^2 - 4x - 1$ and that $2\cos(2\pi/7)$ is a root of $x^3 + x^2 - 2x - 1$. Hence show that a regular septagon is not constructible.

13.18. If the regular n-gon is constructible and $n = qr$, show that the regular q-gon is also constructible.

13.19. Let $\xi = \cos(2\pi/p^2) + i\sin(2\pi/p^2)$. Show that ξ is a root of

$$f(x) = 1 + x^p + x^{2p} + \cdots + x^{(p-1)p}.$$

Prove that $f(x)$ is irreducible over $\mathbb{Q}$ by applying Eisenstein's criterion to $f(1 + x)$.

13.20. Using Exercises 13.18 and 13.19, prove that if a regular n-gon is constructible, then $n = 2^k p_1 \cdots p_r$ where $p_1, \ldots, p_r$ are distinct Fermat primes.

13.21. Prove that a regular 17-gon is constructible.

For Exercises 13.22 to 13.33, can you construct a root of the polynomials?

13.22. $x^2 - 7x - 13$.

13.23. $x^4 - 5$.

13.24. $x^8 - 16$.

13.25. $x^3 - 10x^2 + 2$.

13.26. $x^4 + x^3 - 12x^2 + 7x - 1$.

13.27. $x^5 - 9x^3 + 3$.

13.28. $x^6 + x^3 - 1$.

13.29. $3x^6 - 8x^4 + 1$.

13.30. $4x^4 - x^2 + 2x + 1$.

13.31. $x^4 + x - 1$.

13.32. $x^{48} - 1$.

13.33. $x^4 - 4x^3 + 4x^2 - 2$.

14

ERROR-CORRECTING CODES

With the increased use of electronic instrumentation and computers, there is a growing need for methods of transmitting information quickly and *accurately* over radio and telephone lines and to and from digital storage devices. In fact, CDs and DVDs use error-correcting codes.

Over any transmission line, there is liable to be noise, that is, extraneous signals that can alter the transmitted information. This is not very noticeable in listening to the radio or even in reading a telegram, because normal English is about 20% redundant. However, in transmissions from satellites and in computer link-ups, the redundancy is usually zero; thus we would like to detect, and possibly correct, any errors in the transmitted message. We can do this by putting the message into a code that will detect and correct most of the errors. (These are not the sorts of codes useful to a spy. Secret codes are made deliberately hard to break, whereas error-correcting codes are designed to be easily decoded.)

One familiar code is the parity check digit that is usually attached to each number inside a computer. A number is written in binary form and a check digit is added that is the sum modulo 2 of the other digits. The sum of the digits of any number and its check digit is always even unless an error has occurred. This check digit will detect any odd number of errors but not an even number of errors. This is useful if the probability of two errors occurring in the same word is very small. When a parity check failure occurs in reading words from a computer memory, the computer automatically rereads the faulty word. If a parity check failure occurs inside the arithmetic unit, the program usually has to be rerun.

All the codes we construct are obtained by adding a certain number of check digits to each block of information. Codes can either be used simply to *detect* errors or can be used to *correct* errors. A code that will detect $2t$ or fewer errors can be used to correct t or fewer errors.

Error-detecting codes are used when it is relatively easy to send the original message again, whenever an error is detected. The single parity check code in a computer is an example of an error-detecting code.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

Sometimes it is impossible or too expensive to retransmit the original message when an error is detected. Error-correcting codes then have to be employed. These are used, for example, in transmissions from satellites and space probes (see Figure 14.1). The extra equipment needed to store and retransmit messages from a satellite would add unnecessary weight to the payload. Error-correcting codes are also used when transmitting data from computer memories to storage

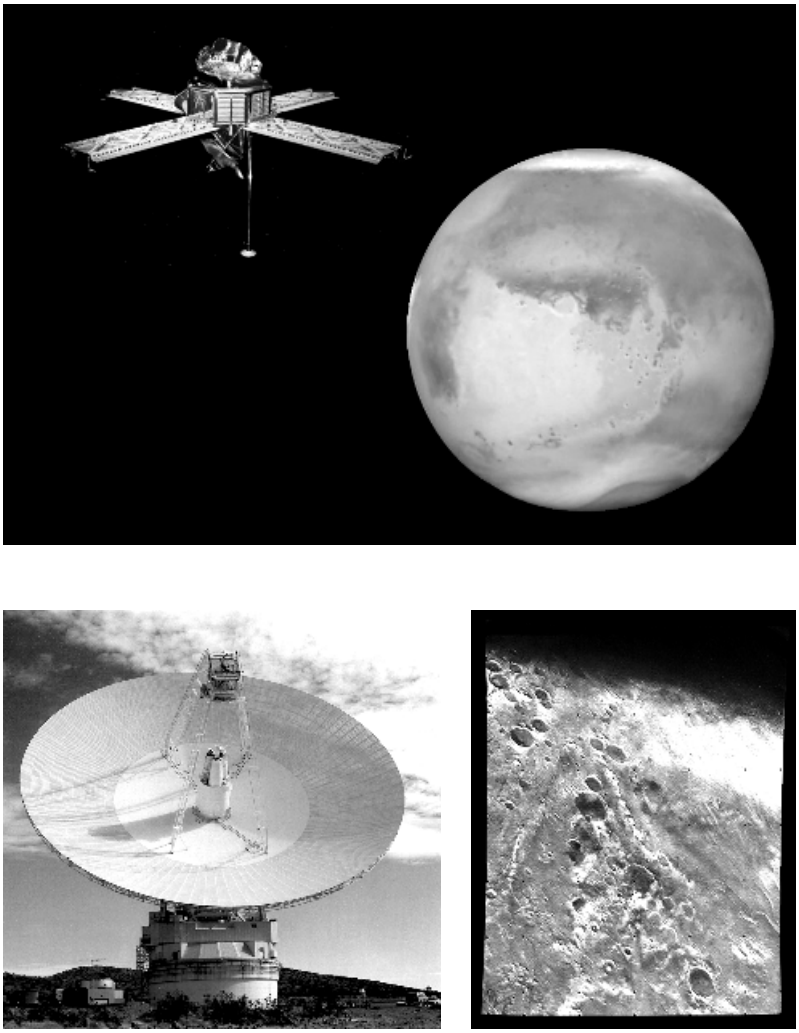


Figure 14.1. In 1969 the *Mariners 6* and 7 space probes sent back over 200 close-up photographs of Mars. Each photograph was divided into 658,240 pixels and each pixel was given a brightness level ranging from 1 to 2^8. Therefore, each photograph required about 5 million bits of information. These bits were encoded, using an error-correcting code, and transmitted at a rate of 16,200 bits per second back to Earth, where they were received and decoded into photographs. (Courtesy of NASA/JPL/Caltech.)

devices. If a message containing an error is stored on a device, it may be weeks before it is read and the error detected; by this time the original data might be lost.

THE CODING PROBLEM

In most digital computers and many communication systems, information is handled in binary form; that is, messages are formed from the symbols 0 and 1. Therefore, in this chapter, we discuss only **binary codes**. However, most of the results generalize to codes whose symbols come from any finite field.

We assume that when a message is transmitted over a channel, the probability of the digit 1 being changed into 0 is the same as that of 0 being changed into 1. Such channels are called **binary symmetric**. Figure 14.2 illustrates what might happen to a message over a noisy channel.

To transmit a message over a noisy channel, we break up the message into blocks of k digits and we **encode** each block by attaching $n - k$ **check digits** to obtain a **code word** consisting of n digits, as shown in Figure 14.3. Such a code is referred to as an (n, k)-**code**.

The code words can now be transmitted over the noisy channel, and after being received, they can be processed in one of two ways. The code can be used to *detect errors* by checking whether or not the received word is a code word. If the received word is a code word, it is assumed to be the transmitted word. If

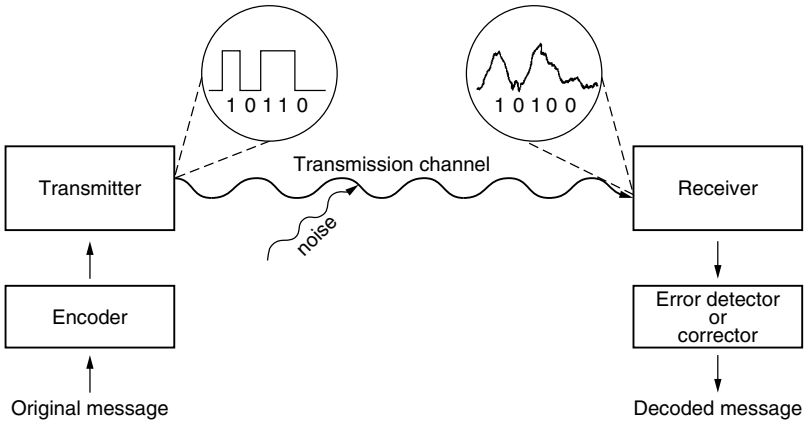


Figure 14.2. Block diagram for error detection or correction.

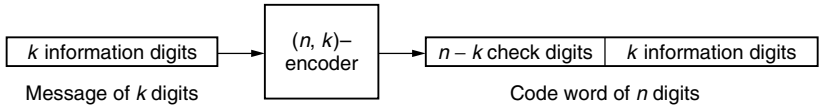


Figure 14.3. Encoding a block of k digits.

the received word is not a code word, an error must have occurred during transmission, and the receiver can request that the word be retransmitted. However, the code could also be used to *correct errors*. In this case, the decoder chooses the transmitted code word that is most likely to produce each received word.

Whether a code is used as an error-detecting or error-correcting code depends on each individual situation. More equipment is required to correct errors, and fewer errors can be corrected than could be detected; on the other hand, when a code only detects errors, there is the trouble of stopping the decoding process and requesting retransmission every time an error occurs.

In an (n, k)-code, the original message is k digits long and there are 2^k different possible messages and hence 2^k code words. The received words have n digits; hence there are 2^n possible words that could be received, only 2^k of which are code words.

The extra $n - k$ check digits that are added to produce the code word are called **redundant digits** because they carry no new information but only allow the existing information to be transmitted more accurately. The ratio $R = k/n$ is called the **code rate** or **information rate**.

For each particular communications channel, it is a major problem to design a code that will transmit useful information as fast as possible and, at the same time, as reliably as possible.

It was proved by C. E. Shannon in 1948 that each channel has a definite capacity C, and for any rate $R < C$, there exist codes of rate R such that the probability of erroneous decoding is arbitrarily small. In other words, by increasing the code length n and keeping the code rate R below the channel capacity C, it is possible to make the probability of erroneous decoding as small as we please. However, this theory provides no useful method of finding such codes.

For codes to be efficient, they usually have to be very long; they may contain 2^{100} messages and many times that number of possible received words. To be able to encode and decode such long codes effectively, we look at codes that have a strong algebraic structure.

SIMPLE CODES

We now compare two very simple codes of length 3. The first is the (3, 2)-code that attaches a single parity check to a message of length 2. The parity check is the sum modulo 2 of the digits in the message. Hence a received word is a code word if and only if it contains an even number of 1's. The code words are given in Table 14.1. The second code is the (3, 1)-code that repeats a message, consisting of a single digit, three times. Its two code words are illustrated in Table 14.2.

If one error occurs in the (3, 2) parity check code during transmission, say 101 is changed to 100, then this would be detected because there would be an odd number of 1's in the received word. However, this code will not correct any errors; the received word 100 is just as likely to have come from 110 or 000 as from 101. This code will not detect two errors either. If 101 was the transmitted

TABLE 14.1. (3, 2) Parity Check Code

Message	Code Word
00	000
01	101
10	110
11	011
	↑
	parity check

TABLE 14.2. (3, 1) Repeating Code

Message	Code Word
0	000
1	111

code word and errors occurred in the first two positions, the received word would be 011, and this would be erroneously decoded as 11.

The decoder first performs a parity check on the received word. If there are an even number of 1's in the word, the word passes the parity check, and the message is the last two digits of the word. If there are an odd number of 1's in the received work, it fails the parity check, and the decoder registers an error. Examples of this decoding are shown in Table 14.3.

The (3, 1) repeating code can be used as an error-detecting code, and it will detect one or two transmission errors but, of course, not three errors. This same code can also be used as an error-correcting code. If the received word contains more 1's than 0's, the decoder assumes that the message is 1; otherwise, it assumes that the message is 0. This will correctly decode messages containing one error, but will erroneously decode messages containing more than one error. Examples of this decoding are shown in Table 14.4.

One useful way to discover the error-detecting and error-correcting capabilities of a code is by means of the Hamming distance. The **Hamming distance** between two words u and v of the same length is defined to be the number of positions in which they differ. This distance is denoted by $d(u, v)$. For example, $d(101, 100) = 1, d(101, 010) = 3$, and $d(010, 010) = 0$.

TABLE 14.3. (3, 2) Parity Check Code Used to Detect Errors

Received Word	101	111	100	000	110
Parity Check	Passes	Fails	Fails	Passes	Passes
Received Message	01	Error	Error	00	10

TABLE 14.4. (3, 1) Repeating Code Used to Correct Errors

Received Word	111	010	011	000
Decoded Message	1	0	1	0

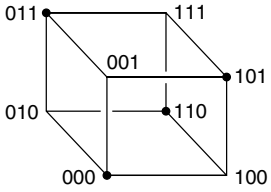


Figure 14.4. The code words of the (3,2) parity check code are shown as large dots.

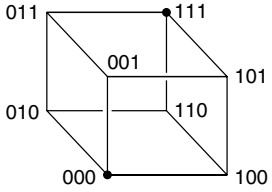


Figure 14.5. The code words of the (3,1) repeating code are shown as large dots.

The Hamming distance between two words is the number of single errors needed to change one word into the other. In an (n, k)-code, the 2^n received words can be thought of as placed at the vertices of an n-dimensional cube with unit sides. The Hamming distance between two words is the shortest distance between their corresponding vertices along the edges of the n-cube. The 2^k code words form a subset of the 2^n vertices, and the code has better error-correcting and error-detecting capabilities the farther apart these code words are. Figure 14.4 illustrates the (3,2) parity check code whose code words are at Hamming distance 2 apart. Figure 14.5 illustrates the (3,1) repeating code whose code words are at Hamming distance 3 apart.

Proposition 14.1. A code will detect all sets of t or fewer errors if and only if the minimum Hamming distance between code words is at least $t + 1$.

Proof. If r errors occur when the code word u is transmitted, the received word v is at Hamming distance r from u. These transmission errors will be detected if and only if v is not another code word. Hence all sets of t or fewer errors in the code word u will be detected if and only if the Hamming distance of u from all the other code words is at least $t + 1$. □

Proposition 14.2. A code is capable of correcting all sets of t or fewer errors if and only if the minimum Hamming distance between code words is at least $2t + 1$.

Proof. Suppose that the code contains two code words u_1 and u_2 at Hamming distance $2t$ or closer. Then there exists a received word v that differs from u_1 and u_2 in t or fewer positions. This received word v could have originated from u_1 or u_2 with t or fewer errors and hence would not be correctly decoded in both these situations.

Conversely, any code whose code words are at least $2t + 1$ apart is capable of correcting up to t errors. This can be achieved in decoding by choosing the code word that is closest to each received word. □

Table 14.5 summarizes these results.

TABLE 14.5. Detection Capabilities of Various Codes

Code	Minimum Distance between Code Words	Number of Errors Detectable	Number of Errors Correctable	Information Rate
(3,2) parity check code	2	1	0	2/3
(3,1) repeating code	3	2	1	1/3
General (n, k) code	d	$d-1$	$\leqslant (d-1)/2$	k/n

POLYNOMIAL REPRESENTATION

There are various ways that a word of n binary digits can be represented algebraically. One convenient way is by means of a polynomial in $\mathbb{Z}_2[x]$ of degree less than n. The word $a_0a_1\cdots a_{n-1}$ can be represented by the polynomial

$$a_0 + a_1x + \cdots + a_{n-1}x^{n-1} \in \mathbb{Z}_2[x].$$

We now use this representation to show how codes can be constructed.

Let $p(x) \in \mathbb{Z}_2[x]$ be a polynomial of degree $n-k$. The **polynomial code generated by** $p(x)$ is an (n, k)-code whose code words are precisely those polynomials, of degree less than n, which are divisible by $p(x)$.

A message of length k is represented by a polynomial $m(x)$, of degree less than k. In order that the higher-order coefficients in a code polynomial carry the message digits, we multiply $m(x)$ by x^{n-k}. This has the effect of shifting the message $n-k$ places to the right. To *encode* the message polynomial $m(x)$, we divide $x^{n-k}m(x)$ by $p(x)$ and add the remainder, $r(x)$, to $x^{n-k}m(x)$ to form the code polynomial

$$v(x) = r(x) + x^{n-k}m(x).$$

This code polynomial is always a multiple of $p(x)$ because, by the division algorithm,

$$x^{n-k}m(x) = q(x)\cdot p(x) + r(x) \quad \text{where} \quad \deg r(x) < n-k \quad \text{or} \quad r(x) = 0;$$

thus

$$v(x) = r(x) + x^{n-k}m(x) = -r(x) + x^{n-k}m(x) = q(x)\cdot p(x).$$

(Remember $r(x) = -r(x)$ in $\mathbb{Z}_2[x]$.) The polynomial $x^{n-k}m(x)$ has zeros in the $n-k$ lowest-order terms, whereas the polynomial $r(x)$ is of degree less than $n-k$; hence the k highest-order coefficients of the code polynomial $v(x)$ are the message digits, and the $n-k$ lowest-order coefficients are the check digits. These check digits are precisely the coefficients of the remainder $r(x)$.

For example, let $p(x) = 1 + x^2 + x^3 + x^4$ be the generator polynomial of a (7, 3)-code. We encode the message 101 as follows:

$$\begin{aligned}
\text{message} &= 1 \quad 0 \quad 1 \\
m(x) &= 1 \quad\quad + x^2 \\
x^4 m(x) &= \quad\quad\quad\quad x^4 \quad + x^6 \\
r(x) &= 1 + x \\
v(x) = r(x) + x^4 m(x) &= 1 + x \quad\quad + x^4 \quad + x^6 \\
\text{code word} &= \underbrace{1 \quad 1 \quad 0 \quad 0}_{\text{check digits}} \quad \underbrace{1 \quad 0 \quad 1}_{\text{message}}
\end{aligned}$$

$$\begin{array}{r|l}
 & x^2 + x + 1 \\
\hline
x^4 + x^3 + x^2 + 0 + 1 & x^6 \quad\quad + x^4 \\
 & x^6 + x^5 + x^4 \quad\quad + x^2 \\
\hline
 & \quad x^5 \quad\quad\quad + x^2 \\
 & \quad x^5 + x^4 + x^3 \quad\quad + x \\
\hline
 & \quad\quad x^4 + x^3 + x^2 + x \\
 & \quad\quad x^4 + x^3 + x^2 \quad\quad + 1 \\
\hline
 & \quad\quad\quad\quad\quad\quad x + 1 \\
\hline
\end{array}$$

The generator polynomial $p(x) = a_0 + a_1 x + \cdots + a_{n-k} x^{n-k}$ is always chosen so that $a_0 = 1$ and $a_{n-k} = 1$, since this avoids wasting check digits. If $a_0 = 0$, any code polynomial would be divisible by x and the first digit of the code word would always be 0; if $a_{n-k} = 0$, the coefficient of x^{n-k-1} in the code polynomial would always be 0.

Example 14.3. Write down all the code words for the code generated by the polynomial $p(x) = 1 + x + x^3$ when the message length k is 3.

Solution. Since deg $p(x) = 3$, there will be three check digits, and since the message length k is 3, the code word length n will be 6. The number of messages is $2^k = 8$.

$$\begin{array}{r|l}
 & x + 1 \\
\hline
x^3 + 0 + x + 1 & x^4 + x^3 \\
 & x^4 \quad\quad + x^2 + x \\
\hline
 & \quad x^3 + x^2 + x \\
 & \quad x^3 \quad\quad\quad x + 1 \\
\hline
 & \quad\quad\quad x^2 \quad + 1 \\
\hline
\end{array}$$

Consider the message 110, which is represented by the polynomial $m(x) = 1 + x$. Its check digits are the coefficients of the remainder $r(x) = 1 + x^2$,

obtained by dividing $x^3m(x) = x^3 + x^4$ by $p(x)$. Hence the code polynomial is $v(x) = r(x) + x^3m(x) = 1 + x^2 + x^3 + x^4$, and the code word is 101110. Table 14.6 shows all the code words. □

A received message can be checked for errors by testing whether it is divisible by the generator polynomial $p(x)$. If the remainder is nonzero when the received polynomial $u(x)$ is divided by $p(x)$, an error must have occurred during transmission. If the remainder is zero, the received polynomial $u(x)$ is a code word, and either no error has occurred or an undetectable error has occurred.

Example 14.4. If the generator polynomial is $p(x) = 1 + x + x^3$, test whether the following received words contain detectable errors: (i) 100011, (ii) 100110, (iii) 101000.

Solution. The received polynomials are $1 + x^4 + x^5$, $1 + x^3 + x^4$, and $1 + x^2$, respectively. These contain detectable errors if and only if they have nonzero remainders when divided by $p(x) = 1 + x + x^3$.

$$
\begin{array}{r|l}
x^3 + x + 1 & x^2 + x + 1 \\
\hline
x^5 + x^4 + 0 + 0 + 0 + 1 & \\
x^5 \quad + x^3 + x^2 & \\
\hline
x^4 + x^3 + x^2 \quad + 1 & \\
x^4 \quad + x^2 + x & \\
\hline
x^3 \quad + x + 1 & \\
x^3 \quad + x + 1 & \\
\hline
0 &
\end{array}
\qquad
\begin{array}{r|l}
x^3 + x + 1 & x + 1 \\
\hline
x^4 + x^3 + 0 + 0 + 1 & \\
x^4 \quad + x^2 + x & \\
\hline
x^3 + x^2 + x + 1 & \\
x^3 \quad + x + 1 & \\
\hline
x^2 &
\end{array}
$$

$$
\begin{array}{r|l}
x^3 + x + 1 & 0 \\
\hline
x^2 + 0 + 1 & \\
0 & \\
\hline
x^2 \quad + 1 &
\end{array}
$$

Hence $1 + x^4 + x^5$ is divisible by $p(x)$, but $1 + x^3 + x^4$ and $1 + x^2$ are not. Therefore, errors have occurred in the latter two words but are unlikely to have occurred in the first. □

Table 14.6 lists all the code words for this code. Hence, in Example 14.4 we can tell at a glance whether a word is a code word simply by noting whether it is on this list. However, in practice, the list of code words is usually so large that it is easier to calculate the remainder when the received polynomial is divided by the generator polynomial.

TABLE 14.6. (6, 3) Code Generated by $1 + x + x^3$

Message			Code Word					
			Check Digits			Message Digits		
0	0	0	0	0	0	0	0	0
1	0	0	1	1	0	1	0	0
0	1	0	0	1	1	0	1	0
0	0	1	1	1	1	0	0	1
1	1	0	1	0	1	1	1	0
1	0	1	0	0	1	1	0	1
0	1	1	1	0	0	0	1	1
1	1	1	0	1	0	1	1	1
↑	↑	↑	↑	↑	↑	↑	↑	↑
1	x	x^2	1	x	x^2	x^3	x^4	x^5

Furthermore, this remainder can easily be computed using shift registers. Figure 14.6 shows a shift register for dividing by $1 + x + x^3$. The square boxes represent unit delays, and the circle with a cross inside denotes a modulo 2 adder (or exclusive OR gate).

The delays are initially zero, and a polynomial $u(x)$ is fed into this shift register with the high-order coefficients first. When all the coefficients of $u(x)$ have been fed in, the delays contain the remainder of $u(x)$ when divided by $1 + x + x^3$. If these are all zero, the polynomial $u(x)$ is a code word; otherwise, a detectable error has occurred. Table 14.7 illustrates this shift register in operation.

The register in Figure 14.6 could be modified to encode messages, because the check digits for $m(x)$ are the coefficients of the remainder when $x^3 m(x)$

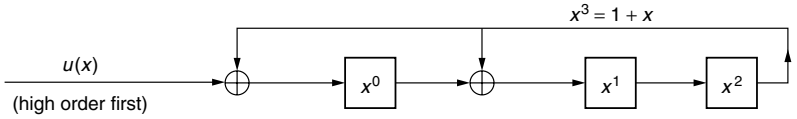


Figure 14.6. Shift register for dividing by $1 + x + x^3$.

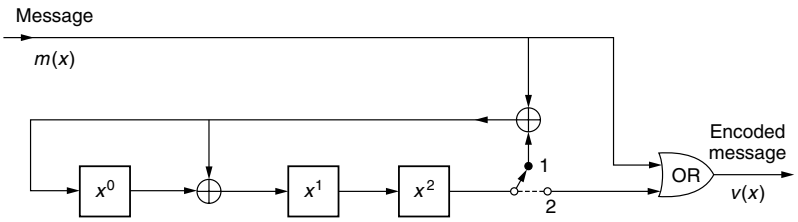


Figure 14.7. Encoding circuit for a code generated by $1 + x + x^3$.

TABLE 14.7. Contents of the Shift Register When $1 + x^3 + x^4$ Is Divided by $1 + x + x^3$

Stage	Received Polynomial Waiting to Enter Register						Register Contents			
							x^0	x^1	x^2	
0	1	0	0	1	1	0	0	0	0	← register initially zero
1		1	0	0	1	1	0	0	0	
2			1	0	0	1	1	0	0	
3				1	0	0	1	1	0	
4					1	0	0	1	1	
5						1	1	1	1	
6							0	0	1	← remainder is x^2

is divided by $1 + x + x^3$. However, the circuit in Figure 14.7 is more efficient for encoding. Here the message $m(x)$ is fed simultaneously to the shift register and the output. While $m(x)$ is being fed in, the switch is in position 1 and the remainder is calculated by the register. Then the switch is changed to position 2, and the check digits are let out to immediately follow the message.

This encoding circuit could also be used for error detection. When $u(x)$ is fed into the encoding circuit with the switch in position 1, the register calculates the remainder of $x^3u(x)$ when divided by $p(x)$. However, $u(x)$ is divisible by $p(x)$ if and only if $x^3u(x)$ is divisible by $p(x)$, assuming that $p(x)$ does not contain a factor x.

How is the generator polynomial chosen so that the code has useful properties without adding too many check digits? We now give some examples.

Proposition 14.5. The polynomial $p(x) = 1 + x$ generates the $(n, n-1)$ parity check code.

Proof. By Proposition 9.33, a polynomial in $\mathbb{Z}_2[x]$ is divisible by $1 + x$ if and only if it contains an even number of nonzero coefficients. Hence the code words of a code generated by $1 + x$ are those words containing an even number of 1's. The check digit for the message polynomial $m(x)$ is the remainder when $xm(x)$ is divided by $1 + x$. Therefore, by the remainder theorem, Theorem 9.4, the check digit is $m(1)$, the parity of the number of 1's in the message. This code is the parity check code. □

The (3, 1) code that repeats the single message digit three times has code words 000 and 111, and is generated by the polynomial $1 + x + x^2$.

We now give one method, using primitive polynomials, of finding a generator for a code that will always detect single, double, or triple errors. Furthermore, the degree of the generator polynomial will be as small as possible so that the check digits are reduced to a minimum. Recall (see Proposition 11.29) that an irreducible polynomial $p(x)$ of degree m over $\mathbb{Z}_2$ is primitive if $p(x)|(1 + x^k)$ for $k = 2^m - 1$ and for no smaller k.

Theorem 14.6. If $p(x)$ is a primitive polynomial of degree m, then the $(n, n-m)$-code generated by $p(x)$ detects all single and double errors whenever $n \leqslant 2^m - 1$.

Proof. Let $v(x)$ be a transmitted code word and $u(x) = v(x) + e(x)$ be the received word. The polynomial $e(x)$ is called the **error polynomial**. An error is detectable if and only if $p(x) \nmid u(x)$. Since $p(x)$ does divide the code word $v(x)$, an error $e(x)$ will be detectable if and only if $p(x) \nmid e(x)$.

If a single error occurs, the error polynomial contains a single term, say x^i, where $0 \leqslant i < n$. Since $p(x)$ is irreducible, it does not have 0 as a root; therefore, $p(x) \nmid x^i$, and the error x^i is detectable.

If a double error occurs, the error polynomial $e(x)$ is of the form $x^i + x^j$ where $0 \leqslant i < j < n$. Hence $e(x) = x^i(1 + x^{j-i})$, where $0 < j - i < n$. Now $p(x) \nmid x^i$, and since $p(x)$ is primitive, $p(x) \nmid (1 + x^{j-i})$ if $j - i < 2^m - 1$. Since $p(x)$ is irreducible, $p(x) \nmid x^i(1 + x^{j-i})$ whenever $n \leqslant 2^m - 1$, and all double errors are detectable. □

Corollary 14.7. If $p_1(x)$ is a primitive polynomial of degree m, the $(n, n-m-1)$-code generated by $p(x) = (1 + x)p_1(x)$ detects all double errors and any odd number of errors whenever $n \leqslant 2^m - 1$.

Proof. The code words in the code generated by $p(x)$ must be divisible by $p_1(x)$ and by $(1 + x)$. The factor $(1 + x)$ has the effect of adding an overall parity check digit to the code. By Proposition 9.33, all the code words have an even number of terms, and the code will detect any odd number of errors. Since the code words are divisible by the primitive polynomial $p_1(x)$, the code will detect all double errors if $n \leqslant 2^m - 1$. □

Some primitive polynomials of low degree are given in Table 14.8. For example, by adding 11 check digits to a message of length 1012 or less, using the generator polynomial $(1 + x)(1 + x^3 + x^{10}) = 1 + x + x^3 + x^4 + x^{10} + x^{11}$, we can detect single, double, triple, and any odd number of errors. Furthermore, the

TABLE 14.8. Short Table of Primitive Polynomials in $\mathbb{Z}_2[x]$

Primitive Polynomial	Degree m	$2^m - 1$
$1 + x$	1	1
$1 + x + x^2$	2	3
$1 + x + x^3$	3	7
$1 + x + x^4$	4	15
$1 + x^2 + x^5$	5	31
$1 + x + x^6$	6	63
$1 + x^3 + x^7$	7	127
$1 + x^2 + x^3 + x^4 + x^8$	8	255
$1 + x^4 + x^9$	9	511
$1 + x^3 + x^{10}$	10	1023

encoding and detecting can be done by a small shift register using only 11 delay units. The number of different messages of length 1012 is 2^{1012}, an enormous figure! When written out in base 10, it would contain 305 digits.

MATRIX REPRESENTATION

Another natural way to represent a word $a_1a_2 \dots a_n$ of length n is by the element $(a_1, a_2, \dots, a_n)^T$ of the vector space $\mathbb{Z}_2^n = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \cdots \times \mathbb{Z}_2$ of dimension n over $\mathbb{Z}_2$. We denote the elements of our vector spaces as column vectors, and $(a_1, a_2, \dots, a_n)^T$ denotes the transpose of $(a_1, a_2, \dots, a_n)$. In an (n, k)-code, the 2^k possible messages of length k are all the elements of the vector space $\mathbb{Z}_2^k$, whereas the 2^n possible received words of length n form the vector space $\mathbb{Z}_2^n$. An encoder is an injective function

$$\gamma: \mathbb{Z}_2^k \to \mathbb{Z}_2^n$$

that assigns to each k digit message an n-digit code word.

An (n, k)-code is called a **linear code** if the encoding function is a linear transformation from $\mathbb{Z}_2^k$ to $\mathbb{Z}_2^n$. Nearly all block codes in use are linear codes, and in particular, all polynomial codes are linear.

Proposition 14.8. Let $p(x)$ be a polynomial of degree $n - k$ that generates an (n, k)-code. Then this code is linear.

Proof. Let $\gamma: \mathbb{Z}_2^k \to \mathbb{Z}_2^n$ be the encoding function defined by the generator polynomial $p(x)$. Let $m_1(x)$ and $m_2(x)$ be two message polynomials of degree less than k and let $\mathbf{m}_1$ and $\mathbf{m}_2$ be the same messages considered as vectors in $\mathbb{Z}_2^k$. The code vector $\gamma(\mathbf{m}_i)$ corresponds to the code polynomial $v_i(x) = r_i(x) + x^{n-k}m_i(x)$, where $r_i(x)$ is the remainder when $x^{n-k}m_i(x)$ is divided by $p(x)$. Now

$$v_1(x) + v_2(x) = r_1(x) + r_2(x) + x^{n-k}[m_1(x) + m_2(x)],$$

and $r_1(x) + r_2(x)$ has degree less than $n - k$; therefore, $r_1(x) + r_2(x)$ is the remainder when $x^{n-k}m_1(x) + x^{n-k}m_2(x)$ is divided by $p(x)$. Hence $v_1(x) + v_2(x)$ corresponds to the code vector $\gamma(\mathbf{m}_1 + \mathbf{m}_2)$ and

$$\gamma(\mathbf{m}_1 + \mathbf{m}_2) = \gamma(\mathbf{m}_1) + \gamma(\mathbf{m}_2).$$

Since the only scalars are 0 and 1, this implies that γ is a linear transformation. □

Let $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ be the standard basis of the vector space $\mathbb{Z}_2^n$, that is, $\mathbf{e}_i$ contains a 1 in the ith position and 0's elsewhere. Let G be the $n \times k$ matrix that represents, with respect to the standard basis, the transformation $\gamma: \mathbb{Z}_2^k \to \mathbb{Z}_2^n$, defined by an (n, k) linear code. This matrix G is called the **generator matrix** or **encoding matrix** of the code.

If $\mathbf{m}$ is a message vector, its code word is $\mathbf{v} = G\mathbf{m}$. The code vectors are the vectors in the image of γ, and they form a vector subspace of $\mathbb{Z}_2^n$ of dimension k. The columns of G are a basis for this subspace, and therefore, a vector is a code vector if and only if it is a linear combination of the columns of the generator matrix G.

(Most coding theorists write the elements of their vector spaces as row vectors instead of column vectors, as used here. In this case, their generator matrix is the transpose of ours, and it operates on the right of the message vector.)

In the (3,2) parity check code, a vector $\mathbf{m} = (m_1, m_2)^T$ is encoded as $\mathbf{v} = (c, m_1, m_2)^T$, where the parity check $c = m_1 + m_2$. Hence the generator matrix is

$$G = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{because} \quad \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} c \\ m_1 \\ m_2 \end{pmatrix}.$$

If the code word is to contain the message digits in its last k positions, the generator matrix must be of the form $G = \begin{pmatrix} P \\ I_k \end{pmatrix}$, where P is an $(n-k) \times k$ matrix and I_k is the $k \times k$ identity matrix.

Example 14.9. Find the generator matrix for the (6,3)-code of Example 14.3 that is generated by the polynomial $1 + x + x^3$.

Solution. The columns of the generator matrix G are the code vectors corresponding to messages consisting of basis elements $\mathbf{e}_1 = (1, 0, 0)^T$, $\mathbf{e}_2 = (0, 1, 0)^T$, and $\mathbf{e}_3 = (0, 0, 1)^T$. We see from Table 14.6 that the generator matrix is

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

□

Any message vector, $\mathbf{m}$, in the (6,3)-code of Example 14.9 can be encoded by calculating $G\mathbf{m}$. However, given any received vector $\mathbf{u}$ it is not easy to determine from the generator matrix G whether or not $\mathbf{u}$ is a code vector. The code vectors form a subspace, Im γ, of dimension k in $\mathbb{Z}_2^n$, generated by the columns of G. We now find a linear transformation $\eta: \mathbb{Z}_2^n \to \mathbb{Z}_2^{n-k}$, represented by a matrix H, whose kernel is precisely Im γ. Hence a vector $\mathbf{u}$ will be a code vector if and only if $H\mathbf{u} = \mathbf{0}$. This proves (ii) in the following theorem.

Theorem 14.10. Let $\gamma: \mathbb{Z}_2^k \to \mathbb{Z}_2^n$ be the encoding function for a linear (n, k)-code with generator matrix $G = \begin{bmatrix} P \\ I_k \end{bmatrix}$, where P is an $(n-k) \times k$ matrix and I_k is the $k \times k$ identity matrix. Then the linear transformation

$$\eta: \mathbb{Z}_2^n \to \mathbb{Z}_2^{n-k}$$

defined by the $(n-k) \times n$ matrix $H = (I_{n-k}|P)$ has the following properties:

(i) Ker $\eta =$ Im γ.
(ii) A received vector $\mathbf{u}$ is a code vector if and only if $H\mathbf{u} = \mathbf{0}$.

Proof. The composition $\eta \circ \gamma : \mathbb{Z}_2^k \to \mathbb{Z}_2^{n-k}$ is the zero transformation because

$$HG = (I_{n-k}|P)\begin{bmatrix} P \\ I_k \end{bmatrix} = (I_{n-k}P + PI_k) = P + P = 0$$

using block multiplication of matrices over the field $\mathbb{Z}_2$. Hence Im $\gamma \subseteq$ Ker η.

Since the first $n-k$ columns of H consist of the standard basis vectors in $\mathbb{Z}_2^{n-k}$, Im η spans $\mathbb{Z}_2^{n-k}$ and contains 2^{n-k} elements. By the morphism theorem for groups,

$$|\text{Ker } \eta| = \frac{|\mathbb{Z}_2^n|}{|\text{Im } \eta|} = \frac{2^n}{2^{n-k}} = 2^k.$$

But Im γ also contains 2^k elements, and therefore Im γ must equal Ker η. □

The $(n-k) \times n$ matrix H in Theorem 14.10 is called the **parity check matrix** of the (n, k)-code.

The parity check matrix of the (3, 2) parity check code is the 1×3 matrix $H = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$. A received vector $\mathbf{u} = (u_1, u_2, u_3)^T$ is a code vector if and only if

$$H\mathbf{u} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = u_1 + u_2 + u_3 = 0.$$

The parity check matrix of the (3, 1)-code that repeats the message three times is the 2×3 matrix $H = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$. A received vector $\mathbf{u} = (u_1, u_2, u_3)^T$ is a code vector if and only if $H\mathbf{u} = \mathbf{0}$, that is, if and only if $u_1 + u_3 = 0$ and $u_2 + u_3 = 0$. In $\mathbb{Z}_2$, this is equivalent to $u_1 = u_2 = u_3$.

The parity check matrix for the (6, 3)-code of Examples 14.3 and 14.9 is

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

The received vector $\mathbf{u} = (u_1, \ldots, u_6)^T$ is a code vector if and only if

$$\begin{array}{ccccccccccc} u_1 & & & + & u_4 & & & + & u_6 & = & 0 \\ & u_2 & & + & u_4 & + & u_5 & + & u_6 & = & 0 \\ & & u_3 & & & + & u_5 & + & u_6 & = & 0. \end{array}$$

That is, if and only if

$$\begin{array}{lllllll} u_1 & = & u_4 & & & + & u_6 \\ u_2 & = & u_4 & + & u_5 & + & u_6 \\ u_3 & = & & & u_5 & + & u_6. \end{array}$$

In this code, the three digits on the right, u_4, u_5, and u_6, are the message digits, whereas u_1, u_2, and u_3 are the check digits. For each code vector $\mathbf{u}$, the equation $H\mathbf{u} = \mathbf{0}$ expresses each check digit in terms of the message digits. This is why H is called the parity check matrix.

Example 14.11. Find the generator matrix and parity check matrix for the (9, 4)-code generated by $p(x) = (1 + x)(1 + x + x^4) = 1 + x^2 + x^4 + x^5$. Then use the parity check matrix to determine whether the word 110110111 is a code word.

Solution. The check digits attached to a message polynomial $m(x)$ are the coefficients of the remainder when $x^5 m(x)$ is divided by $p(x)$. The message polynomials are linear combinations of $1, x, x^2$, and x^3. We can calculate the remainders when x^5, x^6, x^7, and x^8 are divided by $p(x)$ as follows. [This is just like the action of a shift register that divides by $p(x)$.]

$$\begin{aligned} x^5 &\equiv 1 + x^2 + x^4 \bmod p(x) \\ x^6 &\equiv x + x^3 + x^5 \equiv 1 + x + x^2 + x^3 + x^4 \bmod p(x) \\ x^7 &\equiv x + x^2 + x^3 + x^4 + x^5 \equiv 1 + x + x^3 \bmod p(x) \\ x^8 &\equiv x + x^2 + x^4 \bmod p(x). \end{aligned}$$

Therefore, every code polynomial is a linear combination of the following basis polynomials:

$$\begin{array}{lllllllllllll} & & & & 1 & + & x^2 & + & x^4 & + & x^5 & & \\ 1 & + & x & + & x^2 & + & x^3 & + & x^4 & + & x^6 \\ & & & & 1 & + & x & + & x^3 & + & x^7 & & \\ & & & & x & + & x^2 & + & x^4 & + & x^8. & & \end{array}$$

The generator matrix G is obtained from the coefficients of the polynomials above, and the parity check matrix H is obtained from G. Hence

$$G = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

If the received vector is $\mathbf{u} = (1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1)^T$, $H\mathbf{u} = (1 \quad 0 \quad 0 \quad 1 \quad 1)^T$ and hence $\mathbf{u}$ is not a code vector. □

Summing up, if $G = \begin{bmatrix} P \\ \hline I_k \end{bmatrix}$ is the generator matrix of an (n, k)-code, then $H = (I_{n-k}|P)$ is the parity check matrix. We *encode* a message $\mathbf{m}$ by calculating $G\mathbf{m}$, and we can *detect errors* in a received vector $\mathbf{u}$ by calculating $H\mathbf{u}$. A linear code is determined by either giving its generator matrix or by giving its parity check matrix.

ERROR CORRECTING AND DECODING

We would like to find an efficient method for *correcting* errors and decoding. One crude method would be to calculate the Hamming distance between a received word and each code word. The code word closest to the received word would be assumed to be the most likely transmitted word. However, the magnitude of this task becomes enormous as soon as the message length is quite large.

Consider an (n, k) linear code with encoding function $\gamma: \mathbb{Z}_2^k \to \mathbb{Z}_2^n$. Let $V = \text{Im } \gamma$ be the subspace of code vectors. If the code vector $\mathbf{v} \in V$ is sent through a channel and an error $\mathbf{e} \in \mathbb{Z}_2^n$ occurs during transmission, the received vector will be $\mathbf{u} = \mathbf{v} + \mathbf{e}$. The decoder receives the vector $\mathbf{u}$ and has to determine the most likely transmitted code vector $\mathbf{v}$ by finding the most likely error pattern $\mathbf{e}$. This error is $\mathbf{e} = -\mathbf{v} + \mathbf{u} = \mathbf{v} + \mathbf{u}$. The decoder does not know what the code vector $\mathbf{v}$ is, but knows that the error $\mathbf{e}$ lies in the coset $V + \mathbf{u}$.

The most likely error pattern in each coset of $\mathbb{Z}_2^n$ by V is called the **coset leader**.

The coset leader will usually be the element of the coset containing the smallest number of 1's. If two or more error patterns are equally likely, one is chosen arbitrarily. In many transmission channels, errors such as those caused by a stroke of lightning tend to come in bursts that affect several adjacent digits. In these cases, the coset leaders are chosen so that the 1's in each error pattern are bunched together as much as possible.

The cosets of $\mathbb{Z}_2^n$ by the subspace V can be characterized by means of the parity check matrix H. The subspace V is the kernel of the transformation $\eta: \mathbb{Z}_2^n \to \mathbb{Z}_2^{n-k}$; therefore, by the morphism theorem, the set of cosets $\mathbb{Z}_2^n / V$ is isomorphic to Im η, where the isomorphism sends the coset $V + \mathbf{u}$ to $\eta(\mathbf{u}) = H\mathbf{u}$. Hence the coset $V + \mathbf{u}$ is characterized by the vector $H\mathbf{u}$.

If H is an $(n-k) \times n$ parity check matrix and $\mathbf{u} \in \mathbb{Z}_2^n$, then the $(n-k)$-dimensional vector $H\mathbf{u}$ is called the **syndrome** of $\mathbf{u}$. (*Syndrome* is a medical term meaning a pattern of symptoms that characterizes a condition or disease.)

Every element of $\mathbb{Z}_2^{n-k}$ is a syndrome; thus there are 2^{n-k} different cosets and 2^{n-k} different syndromes.

Theorem 14.12. Two vectors are in the same coset of $\mathbb{Z}_2^n$ by V if and only if they have the same syndrome.

Proof. If $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{Z}_2^n$, then the following statements are equivalent:

(i) $V + \mathbf{u}_1 = V + \mathbf{u}_2$, (ii) $\mathbf{u}_1 - \mathbf{u}_2 \in V$,
(iii) $H(\mathbf{u}_1 - \mathbf{u}_2) = 0$, (iv) $H\mathbf{u}_1 = H\mathbf{u}_2$. □

We can decode received words to correct errors by using the following procedure:

1. Calculate the syndrome of the received word.
2. Find the coset leader in the coset corresponding to this syndrome.
3. Subtract the coset leader from the received word to obtain the most likely transmitted word.
4. Drop the check digits to obtain the most likely message.

For a polynomial code generated by $p(x)$, the syndrome of a received polynomial $u(x)$ is the remainder obtained by dividing $u(x)$ by $p(x)$. This is because the jth column of H is the remainder obtained by dividing x^{j-1} by $p(x)$. Hence the syndrome of elements in a polynomial code can easily be calculated by means of a shift register that divides by the generator polynomial.

Example 14.13. Write out the cosets and syndromes for the (6,3)-code with parity check matrix

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

Solution. Each of the rows in Table 14.9 forms a coset with its corresponding syndrome. The top row is the set of code words.

The element in each coset that is most likely to occur as an error pattern is chosen as coset leader and placed at the front of each row. In the top row 000000 is clearly the most likely error pattern to occur. This means that any received word in this row is assumed to contain no errors. In each of the next six rows,

TABLE 14.9. Syndromes and All Words of a (6,3) Code

Syndrome	Coset Leader	Words						
000	000000	110100	011010	111001	101110	001101	100011	010111
100	100000	010100	111010	011001	001110	101101	000011	110111
010	010000	100100	001010	101001	111110	011101	110011	000111
001	001000	111100	010010	110001	100110	000101	101011	011111
110	000100	110000	011110	111101	101010	001001	100111	010011
011	000010	110110	011000	111011	101100	001111	100001	010101
111	000001	110101	011011	111000	101111	001100	100010	010110
101	000110	110010	011100	111111	101000	001011	100101	010001

there is one element containing precisely one nonzero digit; these are chosen as coset leaders. Any received word in one of these rows is assumed to have one error corresponding to the nonzero digit in its coset leader. In the last row, every word contains at least two nonzero digits. We choose 000110 as coset leader. We could have chosen 101000 or 010001, since these also contain two nonzero digits; however, if the errors occur in bursts, then 000110 is a more likely error pattern. Any received word in this last row must contain at least two errors. In decoding with 000110 as coset leader, we are assuming that the two errors occur in the fourth and fifth digits.

Each word in Table 14.9 can be constructed by adding its coset leader to the code word at the top of its column. □

A word could be decoded by looking it up in the table and taking the code word at the top of the column in which it appears. When the code is large, this decoding table is enormous, and it would be impossible to store it in a computer. However, in order to decode, all we really need is the parity check matrix to calculate the syndromes, and the coset leaders corresponding to each syndrome.

Example 14.14. Decode 111001, 011100, 000001, 100011, and 101011 using Table 14.10, which contains the syndromes and coset leaders. The parity check matrix is

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

Solution. Table 14.11 shows the calculation of the syndromes and the decoding of the received words. □

Example 14.15. Calculate the table of coset leaders and syndromes for the (9,4) polynomial code of Example 14.11, which is generated by $p(x) = 1 + x^2 + x^4 + x^5$. □

TABLE 14.10. Syndromes and Coset Leaders for a (6,3) Code

Syndrome	Coset Leader
000	000000
100	100000
010	010000
001	001000
110	000100
011	000010
111	000001
101	000110

TABLE 14.11. Decoding Using Syndromes and Coset Leaders

Word received $\mathbf{u}$	111001	011100	000001	100011	101011
Syndrome $H\mathbf{u}$	000	101	111	000	001
Coset leader $\mathbf{e}$	000000	000110	000001	000000	001000
Code word $\mathbf{u}+\mathbf{e}$	111001	011010	000000	100011	100011
Message	001	010	000	011	011

Solution. There is no simple algorithm for finding all the coset leaders. One method of finding them is as follows.

We write down, in Table 14.12, the 2^5 possible syndromes and try to find their corresponding coset leaders. We start filling in the table by first entering the error patterns, with zero or one errors, next to their syndromes. These will be the most likely errors to occur. The error pattern with one error in the jth position is the jth standard basis vector in $\mathbb{Z}_2^9$ and its syndrome is the jth column of the parity check matrix H, given in Example 14.11. So, for instance, $H(000000001) = 01101$, the last column of H.

The next most likely errors to occur are those with two adjacent errors. We enter all these in the table. For example,

$$\begin{aligned} H(000000011) &= H(000000010) + H(000000001) \\ &= 11010 + 01101, \text{ the last two columns of } H \\ &= 10111. \end{aligned}$$

This still does not fill the table. We now look at each syndrome without a coset leader and find the simplest way the syndrome can be constructed from the columns of H. Most of them come from adding two columns, but some have to be obtained by adding three columns. □

TABLE 14.12. Syndromes and Their Coset Leaders for a (9,4) Code

Syndrome	Coset Leader	Syndrome	Coset Leader	Syndrome	Coset Leader
00000	000000000	01011	000011100	10110	000111000
00001	000010000	01100	011000000	10111	000000011
00010	000100000	01101	000000001	11000	110000000
00011	000110000	01110	011100000	11001	110010000
00100	001000000	01111	000001010	11010	000000010
00101	000000110	10000	100000000	11011	000010010
00110	001100000	10001	001001000	11100	111000000
00111	001110000	10010	000000101	11101	000100100
01000	010000000	10011	001101000	11110	000010100
01001	010010000	10100	000011000	11111	000000100
01010	000001100	10101	000001000		

TABLE 14.13. Decoding Using Syndromes and Coset Leaders

Word received **u**	100110010	100100101	111101100	000111110
Syndrome H**u**	01000	00000	10111	10011
Coset leader **e**	010000000	000000000	000000011	001101000
Code word **u** + **e**	110110010	100100101	111101111	001010110
Message	0010	0101	1111	0110

The (9,4)-code in Example 14.15 will, by Corollary 14.7, *detect* single, double, and triple errors. Hence it will *correct* any single error. It will not detect all errors involving four digits or correct all double errors, because 000000000 and 100001110 are two code words of Hamming distance 4 apart. For example, if the received word is 100001000, whose syndrome is 00101, Table 14.12 would decode this as 100001110 rather than 000000000; both these code words differ from the received word by a double error.

Example 14.16. Decode 100110010, 100100101, 111101100, and 000111110 using the parity check matrix in Example 14.11 and the coset leaders in Table 14.12.

Solution. Table 14.13 illustrates the decoding process. □

BCH CODES

The most powerful class of error-correcting codes known to date were discovered around 1960 by Hocquenghem and independently by Bose and Chaudhuri. For any positive integers m and t, with $t < 2^{m-1}$, there exists a Bose–Chaudhuri–Hocquenghem (BCH) code of length $n = 2^m - 1$ that will correct any combination of t or fewer errors. These codes are polynomial codes with a generator $p(x)$ of degree $\leqslant mt$ and have message length at least $n - mt$.

A **t-error-correcting BCH code** of length $n = 2^m - 1$ has a generator polynomial $p(x)$ that is constructed as follows. Take a primitive element α in the Galois field $GF(2^m)$. Let $p_i(x) \in \mathbb{Z}_2[x]$ be the irreducible polynomial with α^i as a root, and define

$$p(x) = \text{lcm}(p_1(x), p_2(x), \ldots, p_{2t}(x)).$$

It is clear that $\alpha, \alpha^2, \alpha^3, \ldots, \alpha^{2t}$ are all roots of $p(x)$. By Exercise 11.56, $[p_i(x)]^2 = p_i(x^2)$ and hence α^{2i} is a root of $p_i(x)$. Therefore,

$$p(x) = \text{lcm}(p_1(x), p_3(x), \ldots, p_{2t-1}(x)).$$

Since $GF(2^m)$ is a vector space of degree m over $\mathbb{Z}_2$, for any $\beta = \alpha^i$, the elements $1, \beta, \beta^2, \ldots, \beta^m$ are linearly dependent. Hence β satisfies a polynomial of degree at most m in $\mathbb{Z}_2[x]$, and the irreducible polynomial $p_i(x)$ must also

have degree at most m. Therefore,

$$\deg\ p(x) \leqslant \deg\ p_1(x) \cdot \deg\ p_3(x) \cdots \deg\ p_{2t-1}(x) \leqslant mt.$$

Example 14.17. Find the generator polynomials of the t-error-correcting BCH codes of length $n = 15$ for each value of t less than 8.

Solution. Let α be a primitive element of $GF(16)$, where $\alpha^4 + \alpha + 1 = 0$. We repeatedly refer back to the elements of $GF(16)$ given in Table 11.4 when performing arithmetic operations in $GF(16) = \mathbb{Z}_2(\alpha)$.

We first calculate the irreducible polynomials $p_i(x)$ that have α^i as roots. We only need to look at the odd powers of α. The element α itself is the root of $x^4 + x + 1$. Therefore, $p_1(x) = x^4 + x + 1$.

If the polynomial $p_3(x)$ contains α^3 as a root, it also contains

$$(\alpha^3)^2 = \alpha^6, \quad (\alpha^6)^2 = \alpha^{12}, \quad (\alpha^{12})^2 = \alpha^{24} = \alpha^9, \quad \text{and} \quad (\alpha^9)^2 = \alpha^{18} = \alpha^3.$$

Hence

$$\begin{aligned} p_3(x) &= (x - \alpha^3)(x - \alpha^6)(x - \alpha^{12})(x - \alpha^9) \\ &= (x^2 + (\alpha^3 + \alpha^6)x + \alpha^9)(x^2 + (\alpha^{12} + \alpha^9)x + \alpha^{21}) \\ &= (x^2 + \alpha^2 x + \alpha^9)(x^2 + \alpha^8 x + \alpha^6) \\ &= x^4 + (\alpha^2 + \alpha^8)x^3 + (\alpha^9 + \alpha^{10} + \alpha^6)x^2 + (\alpha^{17} + \alpha^8)x + \alpha^{15} \\ &= x^4 + x^3 + x^2 + x + 1. \end{aligned}$$

The polynomial $p_5(x)$ has roots α^5, α^{10}, and $\alpha^{20} = \alpha^5$. Hence

$$\begin{aligned} p_5(x) &= (x - \alpha^5)(x - \alpha^{10}) \\ &= x^2 + x + 1. \end{aligned}$$

The polynomial $p_7(x)$ has roots $\alpha^7, \alpha^{14}, \alpha^{28} = \alpha^{13}, \alpha^{26} = \alpha^{11}$, and $\alpha^{22} = \alpha^7$. Hence

$$\begin{aligned} p_7(x) &= (x - \alpha^7)(x - \alpha^{14})(x - \alpha^{13})(x - \alpha^{11}) \\ &= (x^2 + \alpha x + \alpha^6)(x^2 + \alpha^4 x + \alpha^9) \\ &= x^4 + x^3 + 1. \end{aligned}$$

Now every power of α is a root of one of the polynomials $p_1(x)$, $p_3(x)$, $p_5(x)$, or $p_7(x)$. For example, $p_9(x)$ contains α^9 as a root, and therefore, $p_9(x) = p_3(x)$.

The BCH code that corrects one error is generated by $p(x) = p_1(x) = x^4 + x + 1$.

The BCH code that corrects two errors is generated by

$$p(x) = \text{lcm}(p_1(x), p_3(x)) = (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1).$$

This least common multiple is the product because $p_1(x)$ and $p_3(x)$ are different irreducible polynomials. Hence $p(x) = x^8 + x^7 + x^6 + x^4 + 1$.

The BCH code that corrects three errors is generated by

$$\begin{aligned} p(x) &= \operatorname{lcm}(p_1(x), p_3(x), p_5(x)) \\ &= (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^2 + x + 1) \\ &= x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1. \end{aligned}$$

The BCH code that corrects four errors is generated by

$$\begin{aligned} p(x) &= \operatorname{lcm}(p_1(x), p_3(x), p_5(x), p_7(x)) \\ &= p_1(x) \cdot p_3(x) \cdot p_5(x) \cdot p_7(x) \\ &= \frac{x^{15} + 1}{x + 1} = \sum_{i=0}^{14} x^i. \end{aligned}$$

This polynomial contains all the elements of $GF(16)$ as roots, except for 0 and 1.

Since $p_9(x) = p_3(x)$, the five-error-correcting BCH code is generated by

$$\begin{aligned} p(x) &= \operatorname{lcm}(p_1(x), p_3(x), p_5(x), p_7(x), p_9(x)) \\ &= (x^{15} + 1)/(x + 1), \end{aligned}$$

and this is also the generator of the six- and seven-error-correcting BCH codes. These results are summarized in Table 14.14. □

For example, the two-error-correcting BCH code is a (15, 7)-code with generator polynomial $x^8 + x^7 + x^6 + x^4 + 1$. It contains seven message digits and eight check digits.

The seven-error-correcting code generated by $(x^{15} + 1)/(x + 1)$ has message length 1, and the two code words are the sequence of 15 zeros and the sequence of 15 ones. Each received word can be decoded by majority rule to give the

TABLE 14.14. Construction of t-Error-Correcting BCH Codes of Length 15

t	Roots of $p_{2t-1}(x)$	Degree, $p_{2t-1}(x)$	$p(x)$	$\operatorname{Deg} p(x) = 15 - k$	Message Length, k
1	$\alpha, \alpha^2, \alpha^4, \alpha^8$	4	$p_1(x)$	4	11
2	$\alpha^3, \alpha^6, \alpha^{12}, \alpha^9$	4	$p_1(x)p_3(x)$	8	7
3	α^5, α^{10}	2	$p_1(x)p_3(x)p_5(x)$	10	5
4	$\alpha^7, \alpha^{14}, \alpha^{13}, \alpha^{11}$	4	$(x^{15} + 1)/(x + 1)$	14	1
5	$\alpha^9, \alpha^3, \alpha^6, \alpha^{12}$	4	$(x^{15} + 1)/(x + 1)$	14	1
6	$\alpha^{11}, \alpha^7, \alpha^{14}, \alpha^{13}$	4	$(x^{15} + 1)/(x + 1)$	14	1
7	$\alpha^{13}, \alpha^{11}, \alpha^7, \alpha^{14}$	4	$(x^{15} + 1)/(x + 1)$	14	1

message 1, if the word contains more 1's than 0's, and to give the message 0 otherwise. It is clear that this will correct up to seven errors.

We now show that the BCH code given at the beginning of this section does indeed correct t errors.

Lemma 14.18. The minimum Hamming distance between code words of a linear code is the minimum number of ones in the nonzero code words.

Proof. If v_1 and v_2 are code words, then, since the code is linear, $v_1 - v_2$ is also a code word. The Hamming distance between v_1 and v_2 is equal to the number of 1's in $v_1 - v_2$. The result now follows because the zero word is always a code word, and its Hamming distance from any other word is the number of 1's in that word. □

Theorem 14.19. If $t < 2^{m-1}$, the minimum distance between code words in the BCH code given in at the beginning of this section is at least $2t + 1$, and hence this code corrects t or fewer errors.

Proof. Suppose that the code contains a code polynomial with fewer than $2t + 1$ nonzero terms,

$$v(x) = v_1 x^{r_1} + \cdots + v_{2t} x^{r_{2t}} \quad \text{where} \quad r_1 < \cdots < r_{2t}.$$

This code polynomial is divisible by the generator polynomial $p(x)$ and hence has roots $\alpha, \alpha^2, \alpha^3, \ldots, \alpha^{2t}$. Therefore, if $1 \leqslant i \leqslant 2t$,

$$\begin{aligned} v(\alpha^i) &= v_1 \alpha^{ir_1} + \cdots + v_{2t} \alpha^{ir_{2t}} \\ &= \alpha^{ir_1}(v_1 + \cdots + v_{2t} \alpha^{ir_{2t} - ir_1}). \end{aligned}$$

Put $s_i = r_i - r_1$ so that the elements $v_1, \ldots, v_{2t}$ satisfy the following linear equations:

$$\begin{array}{ccccccccc} v_1 & + & v_2 \alpha^{s_2} & + & \cdots & + & v_{2t} \alpha^{s_{2t}} & = & 0 \\ v_1 & + & v_2 \alpha^{2s_2} & + & \cdots & + & v_{2t} \alpha^{2s_{2t}} & = & 0 \\ \vdots & & & & & \vdots & & & \vdots \\ v_1 & + & v_2 \alpha^{2ts_2} & + & \cdots & + & v_{2t} \alpha^{2ts_{2t}} & = & 0. \end{array}$$

The coefficient matrix is nonsingular because its determinant is the Vandermonde determinant:

$$\det \begin{bmatrix} 1 & \alpha^{s_2} & \cdots & \alpha^{s_{2t}} \\ 1 & \alpha^{2s_2} & & \alpha^{2s_{2t}} \\ \vdots & & & \vdots \\ 1 & \alpha^{2ts_2} & \cdots & \alpha^{2ts_{2t}} \end{bmatrix} = \prod_{2t \geqslant i > j \geqslant 2} (\alpha^{s_i} - \alpha^{s_j}) \neq 0.$$

This determinant is nonzero because $\alpha, \alpha^2, \ldots, \alpha^{2t}$ are all different if $t < 2^{m-1}$. [The expression for the Vandermonde determinant can be verified as follows. When the jth column is subtracted from the ith column, each term contains a factor $(\alpha^{s_i} - \alpha^{s_j})$; hence the determinant contains this factor. Both sides are polynomials in $\alpha^{s_2}, \ldots, \alpha^{s_{2t}}$ of the same degree and hence must differ by a multiplicative constant. By looking at the leading diagonal, we see that this constant is 1.]

The linear equations above must have the unique solution $v_1 = v_2 = \cdots = v_{2t} = 0$. Therefore, there are no nonzero code words with fewer than $2t + 1$ ones, and, by Lemma 14.18 and Proposition 14.2, the code will correct t or fewer errors. □

There is, for example, a BCH (127,92)-code that will correct up to five errors. This code adds 35 check digits to the 92 information digits and hence contains 2^{35} syndromes. It would be impossible to store all these syndromes and their coset leaders in a computer, so decoding has to be done by other methods. The errors in BCH codes can be found by algebraic means without listing the table of syndromes and coset leaders.

In fact, any code with a relatively high information rate must be long and consequently, to be useful, must possess a simple algebraic decoding algorithm. Further details of the BCH and other codes can be found in Roman [46], Lidl and Pilz [10], and Lidl and Niederreiter [34].

EXERCISES

14.1. Which of the following received words contain detectable errors when using the (3, 2) parity check code?

$$110,\ 010,\ 001,\ 111,\ 101,\ 000.$$

14.2. Decode the following words using the (3, 1) repeating code to correct errors:

$$111,\ 011,\ 101,\ 010,\ 000,\ 001.$$

Which of the words contain detectable errors?

14.3. An ancient method of detecting errors when performing the arithmetical operations of addition, multiplication, and subtraction is the method known as **casting out nines**. For each number occurring in a calculation, a check digit is found by adding together the digits in the number and casting out any multiples of nine. The original calculation is then performed on these check digits instead of on the original numbers. The answer obtained, after casting out nines, should equal the check digit of the original answer. If not, an error has occurred. For example, check the following:

$$9642 \times (425 - 163) = 2526204.$$

Add the digits of each number; $9+6+4+2=21=2\times 9+3, 4+2+5=9+2, 1+6+3=9+1$. Cast out the nines and perform the calculation on these check digits:

$$3\times(2-1)=3.$$

Now 3 is the check digit for the answer because $2+5+2+6+2+0+4=2\times 9+3$; hence this calculation checks. Why does this method work?

14.4. Find the redundancy of the English language. Copy a paragraph from a book leaving out every nth letter, and ask a friend to try to read the paragraph. (Try $n=2,3,4,5,6$. If a passage with every fifth letter missing can usually be read, the redundancy is at least $\frac{1}{5}$ or 20%.)

14.5. Each recent book, when published, is given an International Standard Book Number (ISBN) consisting of ten digits, for example, 0-471-29891-3. The first digit is a code for the language group, the second set of digits is a code for the publisher, and the third group is the publisher's number for the book. The last digit is one of $0, 1, 2, \ldots, 9, X$ and is a check digit. Have a look at some recent books and discover how this check digit is calculated. What is the 1×10 parity check matrix? How many errors does this code detect? Will it correct any?

14.6. Is $1+x^3+x^4+x^6+x^7$ or $x+x^2+x^3+x^6$ a code word in the (8,4) polynomial code generated by $p(x)=1+x^2+x^3+x^4$?

14.7. Write down all the code words in the (6,3)-code generated by $p(x)=1+x^2+x^3$.

14.8. Design a code for messages of length 20, by adding as few check digits as possible, that will detect single, double, and triple errors. Also give a shift register encoding circuit for your code.

14.9. Decode the following, using the (6,3)-code given in Table 14.9:

$$000101,\ 011001,\ 110000.$$

14.10. A (7,4) linear code is defined by the equations

$$u_1=u_4+u_5+u_7,\ u_2=u_4+u_6+u_7,\ u_3=u_4+u_5+u_6,$$

where u_4, u_5, u_6, u_7 are the message digits and u_1, u_2, u_3 are the check digits. Write down the generator and parity check matrices for this code. Decode the received words 0000111 and 0001111 to correct any errors.

14.11. Find the minimum Hamming distance between the code words of the code with generator matrix G, where

$$G^T=\begin{bmatrix}0&0&1&0&1&1&0&0&0\\0&1&0&1&0&0&1&0&0\\1&0&1&0&0&0&0&1&0\\0&1&1&0&1&0&0&0&1\end{bmatrix}.$$

Discuss the error-detecting and error-correcting capabilities of this code, and write down the parity check matrix.

14.12. Encode the following messages using the generator matrix of the (9,4)-code of Example 14.11:

$$1101, 0111, 0000, 1000.$$

For Exercises 14.13 to 14.15, find the generator and parity check matrices for the polynomial codes.

14.13. The (4,1)-code generated by $1 + x + x^2 + x^3$.

14.14. The (7,3)-code generated by $(1 + x)(1 + x + x^3)$.

14.15. The (9,4)-code generated by $1 + x^2 + x^5$.

14.16. Find the syndromes of all the received words in the (3,2) parity check code.

14.17. Using the parity check matrix in Example 14.14 and the syndromes in Table 14.10, decode the following words:

$$101110,\ 011000,\ 001011,\ 111111,\ 110011.$$

14.18. Using the parity check matrix in Example 14.11 and the syndromes in Table 14.12, decode the following words:

$$110110110,\ 001001101,\ 111111111,\ 000000111.$$

For Exercises 14.19 to 14.22, construct a table of coset leaders and syndromes for each code.

14.19. The (3,1)-code generated by $1 + x + x^2$.

14.20. The (7,4)-code with parity check matrix

$$H = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

14.21. The (9,4)-code generated by $1 + x^2 + x^5$.

14.22. The (7,3)-code generated by $(1 + x)(1 + x + x^3)$.

14.23. Consider the (63,56)-code generated by $(1 + x)(1 + x + x^6)$.

(a) What is the number of digits in the message before coding?

(b) What is the number of check digits?

(c) How many different syndromes are there?

(d) What is the information rate?

(e) What sort of errors will it detect?

(f) How many errors will it correct?

14.24. One method of encoding a rectangular array of digits is to add a parity check digit to each of the rows and then add a parity check digit to each of the columns (including the column of row checks). For example, in the array in Figure 14.8, the check digits are shaded and the check on checks is crosshatched. This idea is sometimes used when transferring information to and from magnetic tape. The same principle is used in accounting. Show that one error can be corrected and describe how to correct that error. Will it correct two errors? What is the maximum number of errors that it will detect?

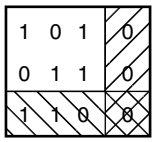


Figure 14.8

14.25. Let V be a vector space over $\mathbb{Z}_p$, where p is a prime. Show that every subgroup is a subspace. Is this result true for a vector space over any Galois field?

14.26. Show that the Hamming distance between vectors has the following properties:
(1) $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$.
(2) $d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w}) \geqslant d(\mathbf{u}, \mathbf{w})$.
(3) $d(\mathbf{u}, \mathbf{v}) \geqslant 0$ with equality if and only if $\mathbf{u} = \mathbf{v}$.
(This shows that d is a **metric** on the vector space.)

14.27. We can use elements from a finite field $GF(q)$, instead of binary digits, to construct codes. If $G = \left[\dfrac{P}{I_k}\right]$ is a generator matrix, show that $H = (I_{n-k} | -P)$ is the parity check matrix.

14.28. Using elements of $\mathbb{Z}_5$, find the parity check matrix of the (7,4)-code generated by $1 + 2x + x^3 \in \mathbb{Z}_5[x]$.

14.29. Find the generators of the two- and three-error-correcting BCH codes of length 15 by starting with the primitive element β in $GF(16)$, where $\beta^4 = 1 + \beta^3$.

14.30. Find the generator polynomial of a single-error-correcting BCH code of length 7.

14.31. Let α be a primitive element in $GF(32)$, where $\alpha^5 = 1 + \alpha^2$. Find an irreducible polynomial in $\mathbb{Z}_2[x]$ with α^3 as a root.

14.32. Find the generator polynomial of a double-error-correcting BCH code of length 31.

14.33. A linear code is called **cyclic** if a cyclic shift of a code word is still a code word; in other words, if $a_1a_2\cdots a_n$ is a code word, then $a_na_1a_2\cdots a_{n-1}$ is also a code word. Show that a binary (n, k) linear code is cyclic if and only if the code words, considered as polynomials, form an ideal in $\mathbb{Z}_2[x]/(x^n - 1)$.

14.34. Let F be a field. Given $f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ in $F[x]$, define the **derivative** $f'(x)$ of $f(x)$ by $f'(x) = a_1 + 2a_2x + \cdots + na_nx^n$. Show that the usual rules of differentiation hold:

(a) $[af(x)]' = af'(x)$.

(b) $[f(x) + g(x)]' = f'(x) + g'(x)$.

(c) $[f(x)g(x)]' = f(x)g'(x) + f'(x)g(x)$.

(d) $\{f[g(x)]\}' = f'[g(x)]g'(x)$.

[*Hint for* (c) *and* (d): Let x and y be two indeterminants over F, and write $F(x, y)$ for the field of fractions of the integral domain $F[x, y]$. Given $f(x)$ in $F[x]$, let $f_0(x, y)$ be the unique polynomial in $F[x, y]$ such that $\dfrac{f(x) - f(y)}{x - y} = f_0(x, y)$ in $F(x, y)$. Show that $f'(x) = f_0(x, x)$.]

14.35. Let a be an element of a field F, and let $f(x) \in F[x]$. Show that $(x - a)^2$ divides $f(x)$ in $F[x]$ if and only if $(x - a)$ divides both $f(x)$ and $f'(x)$. [*Hint*: See Exercise 14.34.]

14.36. In n is odd, show that $x^n - 1$ is square-free when factored into irreducibles in $\mathbb{Z}_2[x]$. [*Hint*: See Exercise 14.35.]

14.37. Write $B_n = \mathbb{Z}_2[x]/(x^n - 1)$ for the factor ring, and write the coset $x + (x^n - 1)$ as $t = x + (x^n - 1)$. Hence binary linear codes are written as follows:

$$\begin{aligned} B_n(t) &= \{a_0 + a_2t + \cdots + a_{n-1}t^{n-1} | a_i \in \mathbb{Z}_2, t^n = 1\} \\ &= \{f(t) | f(x) \in F[x], t^n = 1\}. \end{aligned}$$

Let C denote a cyclic code (see Exercise 14.33).

(a) Show that

$$\begin{aligned} C = (g(t)) &= \{q(t)g(t) | q(t) \in B_n(t)\} \\ &= \{f(t) | g(x) \text{divides} f(x) \text{in } \mathbb{Z}_2[x]\}. \end{aligned}$$

(b) If n is odd, show that $C = (e(t))$ where $[e(t)]^2 = e(t)$ in $B_n(t)$. Show further that $e(t)$ is uniquely determined by C; it is called the **idempotent generator** of C. [*Hint*: By Exercise 14.36 write $x^n - 1 = g(x)h(x)$, where $g(x)$ and $h(x)$ are relatively prime in $\mathbb{Z}_2[x]$.]

Appendix 1

PROOFS

If p and q denote statements, mathematical theorems usually 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 called the **hypothesis**, q is called the **conclusion**, and the verification that $p \Rightarrow q$ is valid is called the **proof** of the implication.

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

Proof. We proceed by assuming that n is odd, and using that information to show that n^2 is also odd. If n is odd it has the form $n = 2k + 1$, where k is some integer. Hence $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1$ is also odd. □

This is called the **direct** method of proof, where the truth of the hypothesis is used directly to establish the truth of the conclusion.

Note that the computation that $n^2 = 2(2k^2 + 2k) + 1$ in Example 1 depends on other properties of arithmetic that we did not prove. In fact, proofs that $p \Rightarrow q$ usually proceed by establishing a sequence $p \Rightarrow p_1 \Rightarrow p_2 \Rightarrow \cdots \Rightarrow p_{n-1} \Rightarrow p_n \Rightarrow q$ of implications leading from p to q. Many of the intervening implications are part of an established part of mathematics, and are not stated explicitly.

Another method is proof by **reduction to cases**. Here is an illustration.

Example 2. Show that $n^2 - n$ is even for every integer n.

Proof. This proposition may not appear to be an implication, but it can be reformulated as: If "n is an integer," then "$n^2 - n$ is even." Given n, the idea is to separate the proof into the two cases that n is even or odd. Since n is even in the first case and $n - 1$ is even in the second case, we see that $n^2 - n = n(n - 1)$ is even in either case. □

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

Note that it is important in Example 2 that every integer n is even or odd, so that the two cases considered cover every possibility. Of course, a proof can proceed by reduction to more than two cases, at least one of which must always hold.

The statements used in mathematics are chosen so that they are either true or false. This leads to another method of proof of an implication $p \Rightarrow q$ called **proof by contradiction**. Since q is either true or false, the idea is to show that it cannot happen that *both* p is true *and* q is false. We accomplish this by showing that the assumption that p is true and q is false leads to a contradiction.

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

Proof. Here we want to prove that $p \Rightarrow q$, where p is the statement that "r is a fraction" and q is the statement that $r^2 \neq 2$. The idea is to show that assuming that p is true and q is false leads to a contradiction. So assume that $r = \dfrac{m}{n}$ is a fraction and $r^2 = 2$. Write $\dfrac{m}{n}$ in lowest terms, so, in particular, m and n are not both even. The statement $r^2 = 2$ leads to $m^2 = 2n^2$, so m^2 is even. Hence m is even (by Example 1), say $m = 2k$, where k is an integer. But then the equation $m^2 = 2n^2$ becomes $4k^2 = 2n^2$, so $n^2 = 2k^2$ is even. Hence n is even (again by Example 1), so we have shown that m and n are both even, contradicting the choice of m and n. This completes the proof. □

As in Example 3, proof by contradiction often provides the simplest verification of an implication.

To provide another example, we need the following concept. An integer greater than 1 is called a **prime** (and we say that it is a **prime number**) if it cannot be factored as the product of two smaller integers both greater than 1. Hence the first few primes are $2, 3, 5, 7, 11, 13, \ldots$, but $6 = 2 \cdot 3$ and $35 = 5 \cdot 7$ are not primes.

Example 4. If $2^n - 1$ is a prime number, show that n is prime.

Proof. We must show that $p \Rightarrow q$ where p is the statement "$2^n - 1$ is prime" and q is the statement that "n is prime." Suppose that q is false, so that $n = ab$ where $a \geqslant 2$ and $b \geqslant 2$ are integers. For convenience, write $k = 2^a$. Then $2^n = 2^{ab} = (2^a)^b = k^b$, and we verify that

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

Since $k \geqslant 4$ this is a factorization of $2^n - 1$ as a product of integers greater than 2, a contradiction. □

The next example illustrates one way to verify 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.

Proof. The first few primes are $n = 2, 3, 5$, and 7, and the corresponding values $2^n - 1 = 3, 7, 31$, and 127 are all prime, as the reader can verify. However, the next prime is $n = 11$, and $2^{11} - 1 = 2047 = 23 \cdot 89$ is not prime. □

We say that $n = 11$ is a **counterexample** to the (proposed) implication in Example 5. Note that it is enough to find even one example in which an implication is not valid to show that the implication is false. Hence it is in a sense easier to disprove an implication than to prove it.

The implications in Examples 4 and 5 are closely related. They have the form $p \Rightarrow q$ and $q \Rightarrow p$, respectively, where p is the statement "$2^n - 1$ is a prime" and q is the statement "n is a prime." In general, each of the statements $p \Rightarrow q$ and $q \Rightarrow p$ is called the **converse** of the other, and these examples show that an implication can be valid even though its converse is not valid.

If both $p \Rightarrow q$ and $q \Rightarrow p$ are valid, we say that p and q are **logically equivalent**. We write this as

$$p \Leftrightarrow q,$$

and read it as "p if and only if q." Many of the most satisfying theorems assert 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."

Proof. The proof that "n is odd" $\Rightarrow$ "n^2 is odd" is given in Example 1. If n^2 is odd, suppose that n is *not* odd. Then n is even, say $n = 2k$, where k is an integer. But then $n^2 = 2(2k)$ is even, a contradiction. Hence the implication $q \Rightarrow p$ has been proved by contradiction. □

Every mathematics book is full of examples of proofs of implications. This is because of the importance of the **axiomatic method**. This procedure arises as follows: In the course of studying various examples, it is observed that they all have certain properties in common. This leads to the study of a general, abstract system where these common properties are *assumed* to hold. The properties are then called **axioms** in the abstract system, and the mathematician proceeds by deducing other properties (called **theorems**) from these axioms using the methods introduced in this appendix. These theorems are then true in all the concrete examples because the axioms hold in each case. The body of theorems is called a **mathematical theory**, and many of the greatest mathematical achievements take this form. Two of the best examples are number theory and group theory, which derive a wealth of theorems from 5 and 4 axioms, respectively.

The axiomatic method is not new: Euclid first used it in about 300 B.C.E. to derive all the propositions of (euclidean) geometry from a list of 10 axioms. His book, *The Elements*, is one of the enduring masterpieces of mathematics.

Appendix 2

INTEGERS

The set $\mathbb{Z} = \{0, \pm 1, \pm 2, \pm 3, \ldots\}$ of **integers** is essential to all of algebra, and has been studied for centuries. In this short section we derive the basic properties of $\mathbb{Z}$, focusing on the idea of the greatest common divisor, and culminating in the prime factorization theorem. We assume a basic knowledge of the addition and multiplication of integers, and of their ordering.

INDUCTION

The following principle is an axiom for the set $\mathbb{P} = \{1, 2, \ldots\}$ of *positive* numbers.

Well-Ordering Axiom. Every nonempty subset of $\mathbb{P}$ has a smallest member.

Our first deduction from the axiom gives a very useful method for proving sequences of statements are true.

Theorem 1. Induction Principle. Let $p_1, p_2, \ldots$ be statements such that:

(i) p_1 is true.
(ii) If p_k is true for some value of $k \geqslant 1$, then p_{k+1} is true.

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

Proof. Let $X = \{n \geqslant 1 | p_n \text{ is false}\}$. If X is nonempty, let m denote the smallest member of X (by the well-ordering axiom). Then $m \neq 1$ by (i), so if we write $n = m - 1$, then $n \geqslant 1$ and p_n is true because $n \notin X$. But then $p_m = p_{n+1}$ is true by (ii), a contradiction. So X is empty, as required. □

Example 2. Prove Gauss' formula: $1 + 2 + \cdots + n = \frac{1}{2}n(n + 1)$ for $n \geqslant 1$.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

Proof. If p_n denotes the statement $1 + 2 + \cdots + n = \frac{1}{2}n(n+1)$ then p_1 is true. If p_k holds, p_{k+1} is true because $1 + 2 + \cdots + k + (k+1) = \frac{1}{2}k(k+1) + (k+1) = \frac{1}{2}(k+1)(k+2)$. Hence every p_k is true by the induction principle. □

There is nothing special about 1 in the induction principle. In fact, the list of statements can be started from any integer b.

Corollary 3. Extended Induction. If b is an integer, let $p_b, p_{b+1}, \ldots$ be statements such that:

(i) p_b is true.
(ii) If p_k is true for some value of $k \geqslant b$, then p_{k+1} is true.

Then p_n is true for every $n \geqslant b$.

Proof. Apply the induction principle to show that the statements $q_1, q_2, \ldots$ are all true, where q_k is the statement that "p_{b+k-1} is true." This means that $p_b, p_{b+1}, \ldots$ are all true, as desired. □

Sometimes it is convenient to be able to replace the inductive assumption that "p_k is true" in Corollary 3 (ii) with the stronger assumption that "each of $p_b, p_{b+1}, \ldots, p_k$ is true." This is valid by the next theorem.

Theorem 4. Strong Induction. If b is an integer, let $p_b, p_{b+1}, \ldots$ be statements such that:

(i) p_b is true.
(ii) If $p_b, p_{b+1}, \ldots, p_k$ are all true for some value of $k \geqslant b$ then p_{k+1} is true.

Then p_n is true for every $n \geqslant b$.

Proof. Apply extended induction to the new statements $q_b, q_{b+1}, \ldots$, where q_k is the statement that "each of $p_b, p_{b+1}, \ldots, p_k$ is true." □

An integer p is called a **prime** if $p \geqslant 2$ and p cannot be written as a product of positive integers apart from $p = 1 \cdot p$. Hence the first few primes are 2, 3, 5, 7, 11, 13, ... With a little experimentation, you can convince yourself that every integer $n \geqslant 2$ is a product of (one or more) primes. Strong induction is needed to prove it.

Theorem 5. Every integer $n \geqslant 2$ is a product of primes.

Proof. Let p_n denote the statement of the theorem. Then p_2 is clearly true. If $p_2, p_3, \ldots, p_k$ are all true, consider the integer $k+1$. If $k+1$ is a prime, there is nothing to prove. Otherwise, $k+1 = ab$, where $2 \leqslant a, b \leqslant k$. But then each of a and b are products of primes because p_a and p_b are both true by the (strong) induction assumption. Hence $ab = k+1$ is also a product of primes, as required. □

Corollary 6. Euclid's Theorem. There are infinitely many primes.

Proof. Suppose on the contrary that $p_1, p_2, \ldots, p_m$ are *all* the primes. In that case, consider the integer $n = 1 + p_1 p_2 \cdots p_m$. It is a product of primes by Theorem 5, so is a multiple of p_i for some $i = 1, 2, \ldots, m$. But then 1 is an integral multiple of p_i, a contradiction. □

Another famous question concerns twin primes, that is, consecutive odd numbers that are both primes: 3 and 5, 5 and 7, 11 and 13, ... The question is whether there are infinitely many twin primes. One curious fact which suggests that there may be only finitely many is that the series $\sum \left\{ \frac{1}{p} \middle| p \text{ a twin prime} \right\}$ of reciprocals of twin primes is convergent, whereas the series $\sum \left\{ \frac{1}{p} \middle| p \text{ a prime} \right\}$ of *all* prime reciprocals is known to be divergent. But the question remains open.

Euclid's theorem certainly implies that there are infinitely many odd primes, that is, primes of the form $2k + 1$, where $k \geqslant 1$. A natural question is whether if a and b are positive integers, there are infinitely many primes of the form $ak + b, k \geqslant 1$. This clearly cannot happen if a and b are both multiples of some integer greater than 1. But it *is* true if 1 is the *only* positive common divisor of a and b, a famous theorem first proved by P. G. L. Dirichlet (1805–1859).

DIVISORS

When we write fractions like $22\frac{1}{7}$ we are using the fact that $22 = 3 \cdot 7 + 1$; that is, when 22 is divided by 7 there is a remainder of 1. The general form of this observation is fundamental to the study of $\mathbb{Z}$.

Theorem 7. Division Algorithm*. Let n and $d \geqslant 1$ be integers. There exist uniquely determined integers q and r such that

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

Proof. Let $X = \{n - td | t \in \mathbb{Z}, n - td \geqslant 0\}$. Then X is nonempty (if $n \geqslant 0$, then $n \in X$; if $n < 0$, then $n(1 - d) \in X$). Hence let r be the smallest member of X (by the well-ordering axiom). Then $r = n - qd$ for some $q \in \mathbb{Z}$, and it remains to show that $r < d$. But if $r \geqslant d$, then $0 \leqslant r - d = n - (q + 1)d$, so $r - d$ is in X contrary to the minimality of r.

As to uniqueness, suppose that $n = q'd + r'$, where $0 \leqslant r' < d$. We may assume that $r \leqslant r'$ (a similar argument works if $r' \leqslant r$). Then $0 \leqslant r' - r = (q' - q)d$, so $(q' - q)d$ is a nonnegative multiple of d that is less than d (because $r' - r \leqslant r' < d$). The only possibility is $(q' - q)d = 0$, so $q' = q$, and hence $r' = r$. □

* This as not an *algorithm* at all, it is a theorem, but the name is well established.

Given n and $d \geqslant 1$, the integers q and r in Theorem 7 are called, respectively, the **quotient** and **remainder** when n is divided by d. For example, if we divide $n = -29$ by $d = 7$, we find that $-29 = (-5) \cdot 7 + 6$, so the quotient is -5 and remainder is 6.

The usual process of long division is a procedure for finding the quotient and remainder for a given n and $d \geqslant 1$. However, they can easily be found with a calculator. For example, if $n = 3196$ and $d = 271$ then $\frac{n}{d} = 11.79$ approximately, so $q = 11$. Then $r = n - qd = 215$, so $3196 = 11 \cdot 271 + 215$, as desired.

If d and n are integers, we say that d **divides** n, or that d is a **divisor** of n, if $n = qd$ for some integer q. We write $d|n$ when this is the case. Thus, a positive integer p is prime if and only if p has no positive divisors except 1 and p. The following properties of the divisibility relation | are easily verified:

(i) $n|n$ *for every* n.
(ii) *If* $d|m$ *and* $m|n$, *then* $d|n$.
(iii) *If* $d|n$ *and* $n|d$, *then* $d = \pm n$.
(iv) *If* $d|n$ *and* $d|m$, *then* $d|(xm + yn)$ *for all integers* x *and* y.

These facts will be used frequently below (usually without comment).

Given positive integers m and n, an integer d is called a **common divisor** of m and n if $d|m$ and $d|n$. The set of common divisors of m and n clearly has a maximum element; what is surprising is that this largest common divisor is actually a *multiple* of every common divisor. With this in mind, we make the following definition: If m and n are integers, not both zero, we say that d is the **greatest common divisor** of m and n, and write $d = \gcd(m, n)$, if the following three conditions are satisfied:

(i) $d \geqslant 1$.
(ii) $d|m$ and $d|n$.
(iii) If $k|m$ and $k|n$, then $k|d$.

In other words, $d = \gcd(m, n)$ is a positive common divisor that is a multiple of every common divisor. It is routine to use conditions (i) to (iii) to show that d is unique if it exists. Note that d does not exist if $m = 0 = n$, but it does exist in every other case (although this is not apparent). In fact, even more is true:

Theorem 8. 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 $X = \{sm + tn | s, t \in \mathbb{Z}; sm + tn \geqslant 1\}$. Then X is not empty since $m^2 + n^2$ is in X, so let d be the smallest member of X (by the well-ordering axiom). Since $d \in X$ we have $d \geqslant 1$ and $d = xm + ym$ for integers x and y, proving conditions (i) and (iii) in the definition of the gcd. Hence it remains to show that $d|m$ and $d|n$. We show that $d|n$; the other is similar. By the division algorithm

write $n = qd + r$, where $0 \leqslant r < d$. Then $r = n - q(xm + yn) = (-qx)m + (1 - qy)n$. Hence, if $r \geqslant 1$, then $r \in X$, contrary to the minimality of d. So $r = 0$ and we have $d|n$. □

When $\gcd(m, n) = xm + yn$ where x and y are integers, we say that $\gcd(m, n)$ is a **linear combination** of m and n. There is an efficient way of computing x and y using the division algorithm. The following example illustrates the method.

Example 9. Find gcd(37, 8) and express it as a linear combination of 37 and 8.

Proof. It is clear that $\gcd(37, 8) = 1$ because 37 is a prime; however, no linear combination is apparent. Dividing 37 by 8, and then dividing each successive divisor by the preceding remainder, gives the first set of equations. The last

$$\begin{aligned} 37 &= 4 \cdot 8 + 5 & 1 &= 3 - 1 \cdot 2 = 3 - 1(5 - 1 \cdot 3) \\ 8 &= 1 \cdot 5 + 3 & &= 2 \cdot 3 - 5 = 2(8 - 1 \cdot 5) - 5 \\ 5 &= 1 \cdot 3 + 2 & &= 2 \cdot 8 - 3 \cdot 5 = 2 \cdot 8 - 3(37 - 4 \cdot 8) \\ 3 &= 1 \cdot 2 + 1 & &= 14 \cdot 8 - 3 \cdot 37 \\ 2 &= 2 \cdot 1 \end{aligned}$$

nonzero remainder is 1, the greatest common divisor, and this turns out always to be the case. Eliminating remainders from the bottom up (as in the second set of equations) gives $1 = 14 \cdot 8 - 3 \cdot 37$. □

The method in Example 9 works in general.

Theorem 10. Euclidean Algorithm. Given integers m and $n \geqslant 1$, use the division algorithm repeatedly:

$$\begin{aligned} m &= q_1 n + r_1 & 0 &\leqslant r_1 < n \\ n &= q_2 r_1 + r_2 & 0 &\leqslant r_2 < r_1 \\ r_1 &= q_3 r_2 + r_3 & 0 &\leqslant r_3 < r_2 \\ &\vdots \qquad \vdots \\ r_{k-2} &= q_k r_{k-1} + r_k & 0 &\leqslant r_k < r_{k-1} \\ r_{k-1} &= q_{k+1} r_k \end{aligned}$$

where in each equation the divisor at the preceding stage is divided by the remainder. These remainders decrease

$$r_1 > r_2 > \cdots \geqslant 0$$

so the process eventually stops when the remainder becomes zero. If $r_1 = 0$, then $\gcd(m, n) = n$. Otherwise, $r_k = \gcd(m, n)$, where r_k is the last nonzero remainder and can be expressed as a linear combination of m and n by eliminating remainders.

Proof. Express r_k as a linear combination of m and n by eliminating remainders in the equations from the second last equation up. Hence every common divisor of m and n divides r_k. But r_k is itself a common divisor of m and n (it divides every r_i—work up through the equations). Hence $r_k = \gcd(m, n)$. □

Two integers m and n are called **relatively prime** if $\gcd(m, n) = 1$. Hence 12 and 35 are relatively prime, but this is not true for 12 and 15 because $\gcd(12, 15) = 3$. Note that 1 is relatively prime to every integer m. The following theorem collects three basic properties of relatively prime integers.

Theorem 11. If m and n are integers, not both zero:

(i) m and n are relatively prime if and only if $1 = xm + yn$ for some integers x and y.
(ii) If $d = \gcd(m, n)$, then $\frac{m}{d}$ and $\frac{n}{d}$ are relatively prime.
(iii) Suppose that m and n are relatively prime.
  (a) If $m|k$ and $n|k$, where $k \in \mathbb{Z}$, then $mn|k$.
  (b) If $m|kn$ for some $k \in \mathbb{Z}$, then $m|k$.

Proof. (i) If $1 = xm + yn$ with $x, y \in \mathbb{Z}$, then every divisor of both m and n divides 1, so must be 1 or -1. It follows that $\gcd(m, n) = 1$. The converse is by the euclidean algorithm.

(ii). By Theorem 8, write $d = xm + yn$, where $x, y \in \mathbb{Z}$. Then $1 = x\frac{m}{d} + y\frac{n}{d}$, and (ii) follows from (i).

(iii). Write $1 = xm + yn$, where $x, y \in \mathbb{Z}$. If $k = am$ and $k = bn, a, b \in \mathbb{Z}$, then $k = kxm + kyn = (xb + ya)mn$, and (a) follows. As to (b), suppose that $kn = qm, q \in \mathbb{Z}$. Then $k = kxm + kyn = (kx + qn)m$, so $m|k$. □

PRIME FACTORIZATION

Recall that an integer p is called a **prime** if:

(i) $p \geqslant 2$.
(ii) *The only positive divisors of* p *are* 1 *and* p.

The reason for not regarding 1 as a prime is that we want the factorization of every integer into primes (as in Theorem 5) to be unique. The following result is needed.

Theorem 12. Euclid's Lemma. Let p denote a prime.

(i) If $p|mn$ where $m, n \in \mathbb{Z}$, then either $p|m$ or $p|n$.

(ii) If $p|m_1 m_2 \cdots m_r$ where each $m_i \in \mathbb{Z}$, then $p|m_i$ for some i.

Proof. (i) Write $d = \gcd(m, p)$. Then $d|p$, so as p is a prime, either $d = p$ or $d = 1$. If $d = p$, then $p|m$; if $d = 1$, then since $p|mn$, we have $p|n$ by Theorem 11.

(ii) This follows from (i) using induction on r. □

By Theorem 5, every integer $n \geqslant 2$ can be written as a product of (one or more) primes. For example, $12 = 2^2 \cdot 3$, $15 = 3 \cdot 5$, $225 = 3^2 \cdot 5^2$. This factorization is unique.

Theorem 13. Prime Factorization Theorem. Every integer $n \geqslant 2$ can be written as a product of (one or more) primes. Moreover, this factorization is unique except for the order of the factors. That is, if

$$n = p_1 p_2 \cdots p_r \quad \text{and} \quad n = q_1 q_2 \cdots q_s,$$

where the p_i and q_j are primes, then $r = s$ and the q_j can be relabeled so that $p_i = q_i$ for each i.

Proof. The existence of such a factorization was shown in Theorem 5. To prove uniqueness, we induct on the minimum of r and s. If this is 1, then n is a prime and the uniqueness follows from Euclid's lemma. Otherwise, $r \geqslant 2$ and $s \geqslant 2$. Since $p_1|n = q_1 q_2 \cdots q_s$ Euclid's lemma shows that p_1 divides some q_j, say $p_1|q_1$ (after possible relabeling of the q_j). But then $p_1 = q_1$ because q_1 is a prime. Hence $\frac{n}{p_1} = p_2 p_3 \cdots p_r = q_2 q_3 \cdots q_s$, so, by induction, $r - 1 = s - 1$ and $q_2, q_3, \ldots, q_s$ can be relabeled such that $p_i = q_i$ for all $i = 2, 3, \ldots, r$. The theorem follows. □

It follows that every integer $n \geqslant 2$ can be written in the form

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

where $p_1, p_2, \ldots, p_r$ are distinct primes, $n_i \geqslant 1$ for each i, and the p_i and n_i are determined uniquely by n. If every $n_i = 1$, we say that n is **square-free**, while if n has only one prime divisor, we call n a **prime power**.

If the prime factorization $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$ of an integer n is given, and if d is a positive divisor of n, then these p_i are the only possible prime divisors of d (by Euclid's lemma). It follows that

Corollary 14. If the prime factorization of n is $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$, then the positive divisors d of n are given as follows:

$$d = p_1^{d_1} p_2^{d_2} \cdots p_r^{d_r} \quad \text{where} \quad 0 \leqslant d_i \leqslant n_i \quad \text{for each } i.$$

This gives another characterization of the greatest common divisor of two positive integers m and n. In fact, let $p_1, p_2, \ldots, p_r$ denote the distinct primes that divide one or the other of m and n. If we allow zero exponents, these numbers can be written in the form

$$n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r} \qquad n_i \geqslant 0$$
$$m = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r} \qquad m_i \geqslant 0.$$

It follows from Corollary 14 that the positive common divisors d of m and n have the form

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

where $0 \leqslant d_i \leqslant \min(m_i, n_i)$ for each i. [Here $\min(m_i, n_i)$ denotes the smaller of the integers m_i and n_i.] Clearly then, we obtain $\gcd(m, n)$ if we set $d_i = \min(m_i, n_i)$ for each i. Before recording this observation (in Theorem 15 below), we first consider a natural question: What if we use $\max(m_i, n_i)$ for each exponent? [Here $\max(m_i, n_i)$ is the larger of the integers m_i and n_i.] This leads to the *dual* of the notion of a greatest common divisor.

If m and n are positive integers, write $n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$ and $m = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r}$ where, as before, the p_i are distinct primes and we have $m_i \geqslant 0$ and $n_i \geqslant 0$ for each i. We define the **least common multiple** of m and n, denoted $\operatorname{lcm}(m, n)$, by

$$\operatorname{lcm}(m, n) = p_1^{\max(m_1, n_1)} p_2^{\max(m_2, n_2)} \cdots p_r^{\max(m_r, n_r)}.$$

It is clear by Corollary 14 that $\operatorname{lcm}(m, n)$ is a common multiple of m and n, and that it is a divisor of any such common multiple. Hence $\operatorname{lcm}(m, n)$ is indeed playing a role dual to that of the greatest common divisor. This discussion is summarized in

Theorem 15. Suppose that m and n are positive integers, and write

$$n = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r} \qquad n_i \geqslant 0$$
$$m = p_1^{m_1} p_2^{m_2} \cdots p_r^{m_r} \qquad m_i \geqslant 0,$$

where the p_i are distinct primes. Then:

$$\gcd(m, n) = p_1^{\min(m_1, n_1)} p_2^{\min(m_2, n_2)} \cdots p_r^{\min(m_r, n_r)}$$
$$\operatorname{lcm}(m, n) = p_1^{\max(m_1, n_1)} p_2^{\max(m_2, n_2)} \cdots p_r^{\max(m_r, n_r)}.$$

The fact that $\max(m, n) + \min(m, n) = m + n$ for any integers m and n gives immediately:

Corollary 16. $mn = \gcd(m, n)\operatorname{lcm}(m, n)$ for all positive integers m and n.

Example 17. Find gcd(600, 294) and lcm(600, 294).

Proof. We have $600 = 2^3 \cdot 3 \cdot 5^2$ and $294 = 3 \cdot 2 \cdot 7^2$ so, as above, write

$$600 = 2^3 3^1 5^2 7^0$$
$$294 = 2^1 3^1 5^0 7^2.$$

Then $\gcd(600, 294) = 2^1 3^1 5^0 7^0 = 6$, while $\operatorname{lcm}(600, 294) = 2^3 3^1 5^2 7^2 = 29{,}400$. Note that Corollary 16 is verified by the fact that $600 \cdot 294 = 6 \cdot 29{,}400$. □

Of course, using Theorem 15 requires finding the prime factorizations of the integers m and n, and that is not easy. One useful observation is that if $n \geqslant 2$ is not a prime, then it has a prime factor $p \leqslant \sqrt{n}$ (it cannot have two factors greater than $\sqrt{n}$), so when looking for prime divisors of n it is only necessary to test the primes $p \leqslant \sqrt{n}$. But for large integers, this is difficult, if not impossible. The euclidean algorithm (and Corollary 16) is a better method for finding greatest common divisors and least common multiples.

Note that this all generalizes: Given a finite collection $a, b, c, \ldots$ of positive integers, write them as

$$\begin{aligned} a &= p_1^{a_1} p_2^{a_2} \cdots p_r^{a_r} \quad a_i \geqslant 0 \\ b &= p_1^{b_1} p_2^{b_2} \cdots p_r^{b_r} \quad b_i \geqslant 0 \\ c &= p_1^{c_1} p_2^{c_2} \cdots p_r^{c_r} \quad c_i \geqslant 0, \\ &\vdots \qquad\qquad\qquad \vdots \end{aligned}$$

where the p_i are the distinct primes that divide at least one of $a, b, c, \ldots$. Then define their **greatest common divisor** and **least common multiple** as follows:

$$\gcd(a, b, c, \ldots) = p_1^{\min(a_1, b_1, c_1, \cdots)} p_2^{\min(a_2, b_2, c_2, \cdots)} \cdots p_r^{\min(a_r, b_r, c_r, \cdots)}$$
$$\operatorname{lcm}(a, b, c, \ldots) = p_1^{\max(a_1, b_1, c_1, \cdots)} p_2^{\max(a_2, b_2, c_2, \cdots)} \cdots p_r^{\max(a_r, b_r, c_r, \cdots)}.$$

Then Theorem 15 extends as follows: $\gcd(a, b, c, \ldots)$ is the common divisor of $a, b, c, \ldots$, that is, a multiple of every such common divisor, and $\operatorname{lcm}(a, b, c, \ldots)$ is the common multiple of $a, b, c, \ldots$, that is, a divisor of every such common multiple.

This is as far as we go into number theory, the study of the integers, a subject that has fascinated mathematicians for centuries. There remain many unanswered questions, among them the celebrated **Goldbach conjecture** that every even number greater than 2 is the sum of two primes. This appears to be very difficult, but it is known that every sufficiently large even number is the sum of a prime and a number that is the product of at most two primes.

However, the twentieth century brought one resounding success. The fact that $3^2 + 4^2 = 5^2$ shows that the equation $a^k + b^k = c^k$ has integer solutions if $k = 2$.

However, Fermat asserted that there are no positive integer solutions if $k \geqslant 3$. He wrote a note in his copy of *Arithmetica* by Diophantus that "I have discovered a truly remarkable proof but the margin is to small to contain it." The result became known as **Fermat's last theorem** and remained open for 300 years. But in 1997, Andrew Wiles proved the result: He related Fermat's conjecture to a problem in geometry, which he solved.

BIBLIOGRAPHY AND REFERENCES

Proofs in Mathematics

1. Bloch, Ethan D., *Proofs and Fundamentals: A First Course in Abstract Mathematics*. Boston: Birkhauser, 2000.
2. Schumacher, Carol, *Chapter Zero: Fundamental Notions of Abstract Mathematics*, 2nd ed. Reading, Mass.: Addison-Wesley, 2000.
3. Solow, Daniel, *How to Read and Do Proofs: An Introduction to Mathematical Thought Processes*, 3rd ed. New York: Wiley, 2002.

Modern Algebra in General

4. Artin, Michael, *Algebra*. Upper Saddle River, N.J.: Prentice Hall, 1991.
5. Birkhoff, Garrett, and Thomas C. Bartee, *Modern Applied Algebra*. New York: McGraw-Hill, 1970.
6. Birkhoff, Garrett, and Saunders Maclane, *A Survey of Modern Algebra*, 4th ed. New York: Macmillan, 1977.
7. Durbin, John R., *Modern Algebra: An Introduction*, 4th ed. New York: Wiley, 2000.
8. Gallian, Joseph A., *Contemporary Abstract Algebra*, 5th ed. Boston: Houghton Mifflin, 2002.
9. Herstein, I. N., *Topics in Algebra*, 2nd ed. New York: Wiley, 1973.
10. Lidl, Rudolf, and Gunter Pilz, *Applied Abstract Algebra*, 2nd ed. New York: Springer-Verlag, 1997.
11. Nicholson, W. Keith, *Introduction to Abstract Algebra*, 2nd ed. New York: Wiley, 1999.
12. Weiss, Edwin, *First Course in Algebra and Number Theory*. San Diego, Calif.: Academic Press, 1971.

History of Modern Algebra

13. Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Vol. 3. New York: Oxford University Press, 1990 (Chap. 49).
14. Stillwell, John, *Mathematics and Its History*, 2nd ed. New York: Springer-Verlag, 2002.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4

Copyright © 2004 John Wiley & Sons, Inc.

Connections to Computer Science and Combinatorics

15. Biggs, Norman L., *Discrete Mathematics*, 2nd ed. Oxford: Oxford University Press, 2003.
16. Davey, B. A., and H. A. Priestley, *Introduction to Lattices and Order*, 2nd ed. Cambridge: Cambridge University Press, 2002.
17. Gathen, Joachim von zur, and Jürgen Gerhard, *Modern Computer Algebra*, 2nd ed. Cambridge: Cambridge University Press, 2003.
18. Hopcroft, John E., Rajeev Motwani, and Jeffrey D. Ullman, *Introduction to Automata Theory, Languages, and Computation*, 2nd ed. Reading, Mass.: Addison-Wesley, 2000.
19. Knuth, Donald E., *The Art of Computer Programming*, Vol. 2, *Seminumerical Algorithms*, 3rd ed. Reading, Mass.: Addison-Wesley, 1998.
20. Kolman, Bernard, Robert C. Busby, and Sharon Cutler Ross, *Discrete Mathematical Structures*, 4th ed. Upper Saddle River, N.J.: Prentice Hall, 1999.
21. Mendelson, Elliott, *Schaum's Outline of Theory and Problems of Boolean Algebra and Switching Circuits*. New York: McGraw-Hill, 1970.
22. Stone, Harold S., *Discrete Mathematical Structures and Their Applications*. Chicago: Science Research Associates, 1973.
23. Whitesitt, J. Eldon, *Boolean Algebra and Its Applications*. New York: Dover, 1995.

Groups and Symmetry

24. Armstrong, Mark Anthony, *Groups and Symmetry*. New York: Springer-Verlag, 1988.
25. Baumslag, Benjamin, and Bruce Chandler, *Schaum's Outline of Group Theory*. New York: McGraw-Hill, 1968.
26. Budden, F. J., *The Fascination of Groups*. Cambridge: Cambridge University Press, 1972.
27. Coxeter, H. S. M., *Introduction to Geometry*, 2nd ed. New York: Wiley, 1989.
28. Cundy, H. Martyn, and A. P. Rollett, *Mathematical Models*, 3rd ed. Stradbroke, Norfolk, England: Tarquin, 1981.
29. Field, Michael, and Martin Golubitsky, *Symmetry in Chaos: A Search for Pattern in Mathematics, Art and Nature*. Oxford: Oxford University Press, 1992.
30. Hall, Marshall, Jr., *The Theory of Groups*. New York: Macmillan, 1959 (reprinted by the American Mathematical Society, 1999).
31. Lomont, John S., *Applications of Finite Groups*. New York: Dover, 1993.
32. Shapiro, Louis W., Finite groups acting on sets with applications. *Mathematics Magazine*, 46 (1973), 136–147.

Rings and Fields

33. Cohn, P. M., *Introduction to Ring Theory*. New York: Springer-Verlag, 2000.
34. Lidl, Rudolf, and Harald Niederreiter, *Introduction to Finite Fields and Their Applications*, rev. ed. Cambridge: Cambridge University Press, 1994.
35. Stewart, Ian, *Galois Theory*, 3rd ed. Boca Raton, Fla.: CRC Press, 2003.

Convolution Fractions

36. Erdelyi, Arthur, *Operational Calculus and Generalized Functions*. New York: Holt, Rinehart and Winston, 1962.
37. Marchand, Jean Paul, *Distributions: An Outline*. Amsterdam: North-Holland, 1962.

Latin Squares

38. Ball, W. W. Rouse, and H. S. M. Coxeter, *Mathematical Recreations and Essays*. New York: Dover, 1987.

39. Lam, C. W. H., The search for a finite projective plane of order 10. *American Mathematical Monthly*, 98(1991), 305–318.
40. Laywine, Charles F., and Gary L. Mullen, *Discrete Mathematics Using Latin Squares*. New York: Wiley, 1998.

Geometrical Constructions

41. Courant, Richard, Herbert Robbins, and Ian Stewart, *What Is Mathematics?* New York: Oxford University Press, 1996.
42. Kalmanson, Kenneth, A familiar constructibility criterion. *American Mathematical Monthly*, 79(1972), 277–278.
43. Kazarinoff, Nicholas D., *Ruler and the Round*. New York: Dover, 2003.
44. Klein, Felix, *Famous Problems of Elementary Geometry*. New York: Dover, 1956.

Coding Theory

45. Kirtland, Joseph, *Identification Numbers and Check Digit Schemes*. Washington, D.C.: Mathematical Association of America, 2001.
46. Roman, Steven, *Introduction to Coding and Information Theory*. New York: Springer-Verlag, 1997.

ANSWERS TO THE ODD-NUMBERED EXERCISES

CHAPTER 2

2.1. Always true.

2.3. When $A \cap (B \Delta C) = \emptyset$.

2.5. When $A \cap (B \Delta C) = \emptyset$.

2.13. $|A \cup B \cup C \cup D| = |A| + |B| + |C| + |D| - |A \cap B|$
$-|A \cap C| - |A \cap D| - |B \cap C| - |B \cap D| - |C \cap D|$
$+|A \cap B \cap C| + |A \cap B \cap D| + |A \cap C \cap D| + |B \cap C \cap D|$
$-|A \cap B \cap C \cap D|$.

2.15. 4.

2.17. Yes; $\mathcal{P}(\emptyset)$.

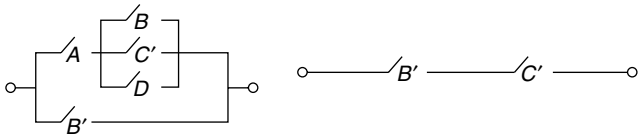
2.25.

A	B	**(a)**	**(b)**	**(c)**	**(d)**
T	T	T	T	T	T
T	F	F	F	T	T
F	T	T	T	F	F
F	F	T	T	T	T

(a) and (b) are equivalent and (c) and (d) are equivalent.

2.27. (a) is a contradiction and (b), (c) and (d) are tautologies.

2.29.



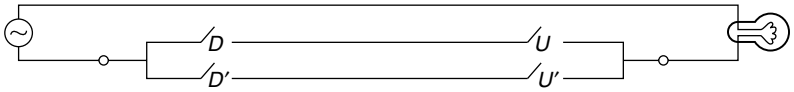
2.31. $B' \wedge C'$.

2.33. $A \vee (B \wedge A')$; $(A \wedge B) \vee (A \wedge B') \vee (A' \wedge B)$; $A \vee B$.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

2.35. $(A \vee B) \wedge (A' \vee B) \wedge (A' \vee B')$; $A' \wedge B$; $A' \wedge B$.

2.39.



2.41. $(A \wedge (B \vee C)) \vee (B \wedge C)$.

2.43. $(A \wedge B) \vee (A' \wedge B') \vee C$.

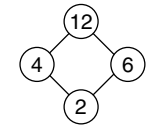
2.45. $A \vee C \vee D$.

2.47. Orange: $(A' \wedge B' \wedge ((C' \wedge D) \vee (C \wedge D'))) \vee (((A' \wedge B) \vee (A \wedge B')) \wedge C' \wedge D')$. Green: $A \wedge B \wedge C \wedge D$.

2.49. Let the result of multiplying AB by CD be EF. Then the circuit for E is $(A' \wedge B \wedge C) \vee (A \wedge ((B \vee C') \wedge D) \vee (B' \wedge C \wedge D'))$, and the circuit for F is $((A \wedge C) \vee (B \wedge D)) \wedge (A' \vee B' \vee C' \vee D')$.

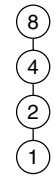
2.51. $(A \vee B) \wedge (A' \vee B')$; $(A \vee B) \wedge (A \vee B') \wedge (A' \vee B')$.

2.53.



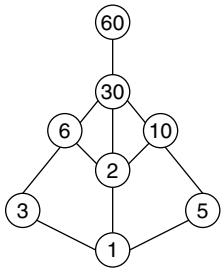
Boolean algebra

2.55.



Lattice

2.57.



Lattice

2.59. The primes p_i.

2.65. Yes.

2.67. $d = a \wedge b' \wedge c'$.

CHAPTER 3

3.1.

$\cdot$	e	g	g^2	g^3	g^4
e	e	g	g^2	g^3	g^4
g	g	g^2	g^3	g^4	e
g^2	g^2	g^3	g^4	e	g
g^3	g^3	g^4	e	g	g^2
g^4	g^4	e	g	g^2	g^3

3.3. See Table 8.3.

3.5. Abelian group.

3.7. Abelian group.

3.9. Not a group; the operation is not closed.

3.11. Abelian group.

3.13. Abelian group.

3.15. Group.

3.25. No.

3.27. D_2.

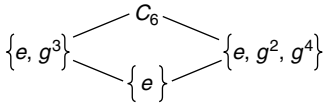
3.29. D_6.

3.31. C_6.

3.33. This is the group O(2) we meet in Chapter 5.

3.35. $\mathbb{Z}$, generated by a glide reflection.

3.37.



3.39. No.

3.41. For any $c \in \mathbb{Q}$, $f: \mathbb{Z} \to \mathbb{Q}$ defined by $f(n) = cn$ for all $n \in \mathbb{Z}$.

3.43. No; $\mathbb{Q}^*$ has an element of order 2, whereas $\mathbb{Z}$ does not.

3.45. The identity has order 1; $(12) \circ (34)$, $(13) \circ (24)$, $(14) \circ (23)$ have order 2, and all the other elements have order 3.

3.47.

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

The identity, 1, has order 1; -1 has order 2; all the other elements have order 4.

3.53. {(1), (123), (132)}.

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

3.57. (12435).

3.59. (165432) is of order 6 and is odd.

3.61. $(1526) \circ (34)$ is of order 4 and is even.

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

3.65. (132).

3.67. {(1), (12), (34), $(12) \circ (34)$, $(13) \circ (24)$, $(14) \circ (23)$, (1324), (1423)}.

3.69. {(1), (13), (24), $(13) \circ (24)$, $(12) \circ (34)$, $(14) \circ (23)$, (1234), (1432)}.

3.73. $\phi(n)$, the number of positive integers less than n that are relatively prime to n.

3.75. 52; 8.

3.77. {e}.

3.83. (1) Achievable; (3) achievable.

3.85. S_3.

3.87. S_2.

3.89. F is not abelian; $y^{-1}x^{-1}y^{-1}xx$.

CHAPTER 4

4.1. Equivalence relation whose equivalence classes are the integers.

4.3. Not an equivalence relation.

4.5.

Left Cosets			Right Cosets		
H	=	$\{(1), (12), (34), (12)\circ(34)\}$	H	=	$\{(1), (12), (34), (12)\circ(34)\}$
$(13)H$	=	$\{(13), (123), (134), (1234)\}$	$H(13)$	=	$\{(13), (132), (143), (1432)\}$
$(14)H$	=	$\{(14), (124), (143), (1243)\}$	$H(14)$	=	$\{(14), (142), (134), (1342)\}$
$(23)H$	=	$\{(23), (132), (234), (1342)\}$	$H(23)$	=	$\{(23), (123), (243), (1243)\}$
$(24)H$	=	$\{(24), (142), (243), (1432)\}$	$H(24)$	=	$\{(24), (124), (234), (1234)\}$
$(1324)H$	=	$\{(1324), (14)\circ(23), (13)\circ(24), (1423)\}$	$H(1324)$	=	$\{(1324), (13)\circ(24), (14)\circ(23), (1423)\}$

4.7. Not a morphism.

4.9. A morphism; $\operatorname{Ker} f = 4\mathbb{Z}$, and $\operatorname{Im} f = \{(0, 0), (1, 1), (0, 2), (1, 3)\}$.

4.11. Not a morphism.

4.19. No.

4.21. $f: C_3 \to C_4$ defined by $f(g^r) = e$.

4.23. $f_k: C_6 \to C_6$ defined by $f_k(g^r) = g^{kr}$ for $k = 0, 1, 2, 3, 4, 5$.

4.25. Not isomorphic; C_{60} contains elements of order 4, whereas $C_{10} \times C_6$ does not.

4.27. Not isomorphic; $C_n \times C_2$ is commutative, whereas D_n is not.

4.29. Not isomorphic; $(1 + i)/\sqrt{2}$ has order 8, whereas $\mathbb{Z}_4 \times \mathbb{Z}_2$ contains no element of order 8.

4.33. C_{10} and D_5.

4.39. $(\mathbb{R}^+, \cdot)$.

4.49. $G_2 \cong S_3$.

4.59. 5 is a generator of $\mathbb{Z}_6^*$, and 3 is a generator of $\mathbb{Z}_{17}^*$.

CHAPTER 5

5.1. C_2 and C_2.

5.3. C_2 and $C_2 \times C_2$.

5.5. C_3 and D_3.

5.7. C_9 and C_9.

5.9. D_4.

5.11. S_4.

5.13. S_4.

5.15. S_4.

5.17. A_5.

5.19. A_5.

5.21. A_5.

5.25. S_4.

5.27. $\begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$.

5.29. D_6.

5.31. C_2.

5.33. D_4 generated by $\begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ and $\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

CHAPTER 6

6.1. 3.

6.3. 38.

6.5. 78.

6.7. 35.

6.9. 333.
6.11. $(n^6 + 3n^4 + 8n^2)/12$.
6.13. 1.
6.15. 96.
6.17. 30.
6.19. 396.
6.21. 126.
6.23. 96.

CHAPTER 7

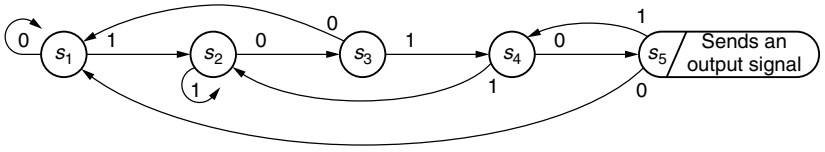
7.1. Monoid with identity 0.
7.3. Semigroup.
7.5. Neither.
7.7. Semigroup.
7.9. Semigroup.
7.11. Monoid with identity 1.
7.13. Neither.
7.15.

gcd	1	2	3	4
1	1	1	1	1
2	1	2	1	2
3	1	1	3	1
4	1	2	1	4

7.17.

$\cdot$	e	c	c^2	c^3	c^4
e	e	c	c^2	c^3	c^4
c	c	c^2	c^3	c^4	c^2
c^2	c^2	c^3	c^4	c^2	c^3
c^3	c^3	c^4	c^2	c^3	c^4
c^4	c^4	c^2	c^3	c^4	c^2

7.19. No; $01 \neq 1$ in the free semigroup.
7.29.



7.31. A congruence relation with quotient semigroup $= \{2\mathbb{N}, 2\mathbb{N} + 1\}$.
7.33. Not a congruence relation.
7.35.

$\star$	[0]	[1]	[00]	[10]	[01]	[010]
[0]	[00]	[01]	[0]	[010]	[1]	[10]
[1]	[10]	[1]	[1]	[10]	[01]	[010]
[00]	[0]	[1]	[00]	[10]	[01]	[010]
[10]	[1]	[01]	[10]	[010]	[1]	[10]
[01]	[010]	[01]	[01]	[010]	[1]	[10]
[010]	[01]	[1]	[010]	[10]	[01]	[010]

7.37. 24.

7.39.

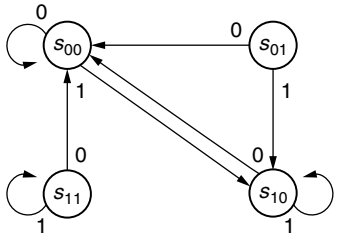
$\star$	$[\Lambda]$	$[\alpha]$	$[\beta]$	$[\gamma]$	$[\alpha\beta]$	$[\alpha\gamma]$
$[\Lambda]$	$[\Lambda]$	$[\alpha]$	$[\beta]$	$[\gamma]$	$[\alpha\beta]$	$[\alpha\gamma]$
$[\alpha]$	$[\alpha]$	$[\Lambda]$	$[\alpha\beta]$	$[\alpha\gamma]$	$[\beta]$	$[\gamma]$
$[\beta]$	$[\beta]$	$[\beta]$	$[\beta]$	$[\gamma]$	$[\beta]$	$[\gamma]$
$[\gamma]$	$[\gamma]$	$[\gamma]$	$[\gamma]$	$[\beta]$	$[\gamma]$	$[\beta]$
$[\alpha\beta]$	$[\alpha\beta]$	$[\alpha\beta]$	$[\alpha\beta]$	$[\alpha\gamma]$	$[\alpha\beta]$	$[\alpha\gamma]$
$[\alpha\gamma]$	$[\alpha\gamma]$	$[\alpha\gamma]$	$[\alpha\gamma]$	$[\alpha\beta]$	$[\alpha\gamma]$	$[\alpha\beta]$

7.41.



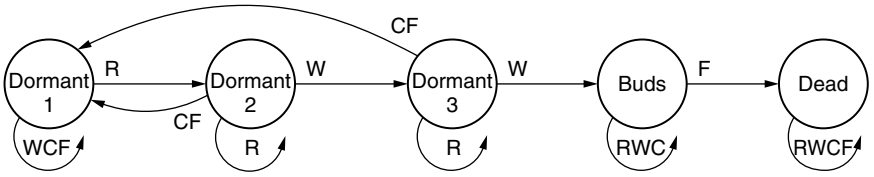
{[0], [1]}.

7.43.



{[0], [1], [10]}.

7.45. The monoid contains 27 elements.



CHAPTER 8

8.1.

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

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

8.3. A ring.

8.5. Not a ring; not closed under multiplication.

8.7. Not a ring; not closed under addition.

8.9. A ring.

8.11. Not a ring; distributive laws do not hold.

8.17. A subring.

8.19. Not a subring; not closed under addition.

8.21. Neither.

8.23. Both.

8.25. Integral domain.

8.29. [2], [4], [5], [6], [8].

8.31. Any nonempty proper subset of X.

8.33. Nonzero matrices with zero determinant.

8.37. $f(x) = [x]_6$.

8.39. $f(x, y) = (x, y)$, (y, x), (x, x) or (y, y).

8.47. (b) -1 and 0.

8.55. The identity is $D_n(x) = (1/2\pi) + (1/\pi)(\cos x + \cos 2x + \cdots + \cos nx)$. The ring is not an integral domain.

CHAPTER 9

9.1. $\frac{3}{2}x^2 + \frac{5}{4}x - \frac{15}{8}$ and $\frac{45}{8}x + \frac{7}{8}$.

9.3. $x^4 + x^3 + x^2 + x$ and 1.

9.5. $3 - i$ and $4 + 2i$, or $4 - i$ and $1 - 2i$, or $4 - 2i$ and $-3 + i$.

9.7. $\gcd(a, b) = 3, s = -5, t = 4$.

9.9. $\gcd(a, b) = 1, s = -(2x + 1)/3, t = (2x + 2)/3$.

9.11. $\gcd(a, b) = 2x + 1, s = 1, t = 2x + 1$.

9.13. $\gcd(a, b) = 1, s = 1, t = -1 + 2i$.

9.15. $x = -6, y = 5$.

9.17. $x = -14, y = 5$.

9.19. [23].

9.21. [17].

9.23. No solutions.

9.25. $(x - 1)(x^4 + x^3 + x^2 + x + 1)$.

9.27. $(x^2 + 2)(x^2 + 3)$.

9.29. $x^4 - 9x + 3$.

9.31. $x^3 - 4x + 1$.

9.33. $(x - \sqrt{2})(x + \sqrt{2})(x - i\sqrt{2})(x + i\sqrt{2})(x - 1 - i)(x - 1 + i)$ $(x + 1 - i)(x + 1 + i)$.

9.35. $(x^2 - 2)(x^2 + 2)(x^2 - 2x + 2)(x^2 + 2x + 2)$.

9.37. $x^5 + x^3 + 1$, $x^5 + x^2 + 1$, $x^5 + x^4 + x^3 + x^2 + 1$, $x^5 + x^4 + x^3 + x + 1$, $x^5 + x^4 + x^2 + x + 1$, $x^5 + x^3 + x^2 + x + 1$.

9.39. $x^3 + 2$.

9.41. $\mathrm{Ker}\psi = \{q(x) \cdot (x^2 - 2x + 4) | q(x) \in \mathbb{Q}[x]\}$ and $\mathrm{Im}\,\psi = \mathbb{Q}(\sqrt{3}i) = \{a + b\sqrt{3}i | a, b \in \mathbb{Q}\}$.

9.43. Irreducible by Eisenstein's Criterion.

9.45. Irreducible, since it has no linear factors.

9.47. Reducible; any polynomial of degree >2 in $\mathbb{R}[x]$ is reducible.

9.49. No.

9.55. No.

9.61. $x \equiv 40 \bmod 42$.

9.63. $x \equiv 22 \bmod 30$.

9.67. 65.

CHAPTER 10

10.1. ((0, 0)), ((0, 1)), ((1, 0)), $\mathbb{Z}_2 \times \mathbb{Z}_2$.

10.3. (0) and $\mathbb{Q}$.

10.5. $(p(x))$ where $p(x) \in \mathbb{C}[x]$.

10.7. The quotient ring is a field.

$+$	(3)	$(3)+1$	$(3)+2$
(3)	(3)	$(3)+1$	$(3)+2$
$(3)+1$	$(3)+1$	$(3)+2$	(3)
$(3)+2$	$(3)+2$	(3)	$(3)+1$

$\cdot$	(3)	$(3)+1$	$(3)+2$
(3)	(3)	(3)	(3)
$(3)+1$	(3)	$(3)+1$	$(3)+2$
$(3)+2$	(3)	$(3)+2$	$(3)+1$

10.9. The ideal $((1, 2))$ is the whole ring $\mathbb{Z}_3 \times \mathbb{Z}_3$. The quotient ring is not a field.

$+$	$((1, 2))$
$((1, 2))$	$((1, 2))$

$\cdot$	$((1, 2))$
$((1, 2))$	$((1, 2))$

10.11. $8x + 2$ and $14x + 97$.

10.13. $x^2 + x$ and x^2.

10.17. (a) 6; (b) 36; (c) $x^2 - 1$, $(a) \cap (b) = (\text{lcm}(a, b))$.

10.33. No.

10.35. The whole ring.

10.37.

$\mathbb{Z}_8$
|
$([2]_8)$
|
$([4]_8)$
|
$([0]_8)$

10.39. Irreducible; $\mathbb{Z}_{11}$.

10.41. Reducible.

10.43. Irreducible; $\mathbb{Q}(\sqrt[4]{2})$.

10.45. Irreducible; $\mathbb{Q}(\sqrt{2}, \sqrt{3})$.

10.47. Not a field; contains zero divisors.

10.49. A field by Corollary 10.16.

10.51. A field by Theorem 10.17.

10.53. Not a field; $x^2 + 1 = (x + 2)(x + 3)$ in $\mathbb{Z}_5[x]$.

10.55. A field isomorphic to $\mathbb{Q}[x]/(x^4 - 11)$.

10.59. (0) and (x^n) for $n \geqslant 0$; (x) is maximal.

CHAPTER 11

11.1. $GF(5) = \mathbb{Z}_5 = \{0, 1, 2, 3, 4\}$.

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

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

11.3. $\mathrm{GF}(9) = \mathbb{Z}_3[x]/(x^2 + 1) = \{a\alpha + b | a, b \in \mathbb{Z}_3, \alpha^2 + 1 = 0\}$.

+	0	1	2	α	$\alpha+1$	$\alpha+2$	2α	$2\alpha+1$	$2\alpha+2$
0	0	1	2	α	$\alpha+1$	$\alpha+2$	2α	$2\alpha+1$	$2\alpha+2$
1	1	2	0	$\alpha+1$	$\alpha+2$	α	$2\alpha+1$	$2\alpha+2$	2α
2	2	0	1	$\alpha+2$	α	$\alpha+1$	$2\alpha+2$	2α	$2\alpha+1$
α	α	$\alpha+1$	$\alpha+2$	2α	$2\alpha+1$	$2\alpha+2$	0	1	2
$\alpha+1$	$\alpha+1$	$\alpha+2$	α	$2\alpha+1$	$2\alpha+2$	2α	1	2	0
$\alpha+2$	$\alpha+2$	α	$\alpha+1$	$2\alpha+2$	2α	$2\alpha+1$	2	0	1
2α	2α	$2\alpha+1$	$2\alpha+2$	0	1	2	α	$\alpha+1$	$\alpha+2$
$2\alpha+1$	$2\alpha+1$	$2\alpha+2$	2α	1	2	0	$\alpha+1$	$\alpha+2$	α
$2\alpha+2$	$2\alpha+2$	2α	$2\alpha+1$	2	0	1	$\alpha+2$	α	$\alpha+1$

·	0	1	2	α	$\alpha+1$	$\alpha+2$	2α	$2\alpha+1$	$2\alpha+2$
0	0	0	0	0	0	0	0	0	0
1	0	1	2	α	$\alpha+1$	$\alpha+2$	2α	$2\alpha+1$	$2\alpha+2$
2	0	2	1	2α	$2\alpha+2$	$2\alpha+1$	α	$\alpha+2$	$\alpha+1$
α	0	α	2α	2	$\alpha+2$	$2\alpha+2$	1	$\alpha+1$	$2\alpha+1$
$\alpha+1$	0	$\alpha+1$	$2\alpha+2$	$\alpha+2$	2α	1	$2\alpha+1$	2	α
$\alpha+2$	0	$\alpha+2$	$2\alpha+1$	$2\alpha+2$	1	α	$\alpha+1$	2α	2
2α	0	2α	α	1	$2\alpha+1$	$\alpha+1$	2	$2\alpha+2$	$\alpha+2$
$2\alpha+1$	0	$2\alpha+1$	$\alpha+2$	$\alpha+1$	2	2α	$2\alpha+2$	α	1
$2\alpha+2$	0	$2\alpha+2$	$\alpha+1$	$2\alpha+1$	α	2	$\alpha+2$	1	2α

11.5. $x^3 + x + 4$.
11.7. Impossible.
11.9. $x^2 + 2$.
11.11. $x^4 - 16x^2 + 16$.
11.13. $8x^6 - 9$.
11.17. 3.

11.19. 2.
11.21. 2.
11.23. ∞.
11.25. ∞.
11.27. $(1 - \sqrt[3]{2} + \sqrt[3]{4})/3$.
11.29. $-(1 + 6\omega)/31$.
11.31. $\alpha^4 + \alpha$.
11.33. 2; not a field.
11.35. 7; a field
11.37. 0; a field.
11.39. 0; not a field.
11.43. $m = 2, 4, p^r$ or $2p^r$, where p is an odd prime; see Weiss [12, Th. 4–6–10].
11.47. $\alpha = \sqrt{2} + \sqrt{-3}$.
11.49. All elements of $GF(32)$ except 0 and 1 are primitive.
11.51. $x^3 + x + 1$.
11.57. No solutions.
11.59. $x = 1$ or $\alpha + 1$.
11.63. 5.
11.65. The output has cycle length 7 and repeats the sequence 1101001, starting at the right.

CHAPTER 12

12.1.

a	b	c	d	e	f	g
b	c	d	e	f	g	a
c	d	e	f	g	a	b
d	e	f	g	a	b	c
e	f	g	a	b	c	d
f	g	a	b	c	d	e
g	a	b	c	d	e	f

12.3. Use $GF(8) = \{0, 1, \alpha, 1 + \alpha, \alpha^2, 1 + \alpha^2, \alpha + \alpha^2, 1 + \alpha + \alpha^2\}$ where $\alpha^3 = \alpha + 1$.
12.7. No.
12.9.

		Week			
		1	2	3	4
		↓	↓	↓	↓
	→	A	B	C	D
Shelf	→	B	C	D	A
height	→	C	D	A	B
	→	D	A	B	C

A, B, C, and D are the four brands of cereal.

12.11.

		Week				
		1 ↓	2 ↓	3 ↓	4 ↓	5 ↓
M	→	*A*	*B*	*C*	*D*	0
T	→	*B*	*C*	*D*	0	*A*
W	→	*C*	*D*	0	*A*	*B*
T	→	*D*	0	*A*	*B*	*C*
F	→	0	*A*	*B*	*C*	*D*

A, *B*, *C*, and *D* are the four different types of music, and 0 refers to no music.

12.15. $y = \alpha x$.

12.17. $y = (\alpha + 2)x + 2\alpha$.

12.21. 1.

12.23.

1	6	11	16
12	15	2	5
14	9	8	3
7	4	13	10

1	6	11	16
15	12	5	2
8	3	14	9
10	13	4	7

12.25.

1	10	19	28	37	46	55	64
35	44	49	58	7	16	21	30
29	22	15	8	57	50	43	36
63	56	45	38	27	20	9	2
52	59	34	41	24	31	6	13
18	25	4	11	54	61	40	47
48	39	62	53	12	3	26	17
14	5	32	23	42	33	60	51

CHAPTER 13

13.1. Constructible.

13.3. Constructible.

13.5. Not constructible.

13.7. No.

13.11. Yes.

13.13. No.

13.15. Yes; $\dfrac{\pi}{21} = \dfrac{1}{4}\left(\dfrac{\pi}{3} - \dfrac{\pi}{7}\right)$.

13.23. Yes.

13.25. No.

13.27. No.

13.29. Yes.

13.31. No.

13.33. Yes.

CHAPTER 14

14.1. 010, 001, 111.

14.3. The checking is done modulo 9, using the fact that any integer is congruent to the sum of its digits modulo 9.

14.5. (1 2 3 4 5 6 7 8 9 10) modulo 11. It will detect one error but not correct any.

14.7. 000000, 110001, 111010, 001011, 101100, 011101, 010110, 100111.

14.9. 101, 001, 100.

14.11. Minimum distance = 3. It detects two errors and corrects one error.

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

14.13. $G^T = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}$, $H = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$.

14.15. $G^T = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$,

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

14.17. 110, 010, 101, 001, 011.

14.19.

Syndrome	Coset Leader
00	000
01	010
10	100
11	001

14.21.

Syndrome	Coset Leader	Syndrome	Coset Leader	Syndrome	Coset Leader
00000	000000000	01011	000010100	10110	000000001
00001	000010000	01100	011000000	10111	000010001
00010	000100000	01101	010000010	11000	110000000
00011	000110000	01110	001000100	11001	000000111
00100	001000000	01111	000000110	11010	100000100
00101	000000010	10000	100000000	11011	000001110
00110	001100000	10001	000001010	11100	000000101
00111	000100010	10010	100100000	11101	000010101
01000	010000000	10011	000000011	11110	000001100
01001	010010000	10100	000001000	11111	000011100
01010	000000100	10101	000011000		

14.23. (**a**) 56; (**b**) 7; (**c**) $2^7 = 128$, (**d**) 8/9; (**e**) it will detect single, double, triple, and any odd number of errors; (**f**) 1.

14.25. No.

14.29. $x^8 + x^4 + x^2 + x + 1$ and $x^{10} + x^9 + x^8 + x^6 + x^5 + x^2 + 1$.

14.31. $x^5 + x^4 + x^3 + x^2 + 1$.

INDEX

Abel, N. H., 1, 47, 48
Abelian group, 48
  finite, 92
Absorption laws, 15
Abstract algebra, 2
Action of a group, 96
Adder, full, 35
  half, 34
  modulo 2, 28
  serial, 152
Addition modulo *m*, 84
Adjoining an element, 220
Affine plane, 242
Algebra, abstract, 2
  boolean, 7, 14, 25
  classical, 1
  modern, 2
Algebraic numbers, 221, 233
Algebraic structure, 4
Algebra of sets, 7
Alternating group, 70, 81, 85, 88
AND gate, 36
Angle trisection, 251, 257
Antisymmetry, 23
Archimedean solids, 121
Archimedes, 258
Associativity, 3, 14, 48, 137, 155
Atom, 26
Automorphism, 75, 151
  Frobenius, 235
Axiomatic method, 295
Axioms, 295
Axis of symmetry, 51

BCH code, 284
Benzene, 126
Biconditional, 18
Bijective, 50, 63
Binary code, 266
Binary operation, 2
Binary symmetric channel, 266
Binomial theorem, 178
  coefficients, 129
Boole, G., 7
Boolean algebra, 7, 14, 25
  isomorphism, 39
  morphism, 39
  representation theorem, 39
Boolean expression, 26
Boolean ring, 158, 176
Bose, R. C., 242, 284
Bridge circuit, 43
Burnside, W., 125
Burnside's theorem (lemma), 125

Cancellation, 53
Cantor's theorem, 41
Carrol, L., 41
Casting out nines, 288
Cauchy sequences, 6
Cayley, A., 47, 71
Cayley's theorem, 71
CD, 264
Center of a group, 74
Characteristic of a ring, 226
Check digit, 266
Chinese remainder theorem, 198
Circle group, 88, 98
Circle squaring, 251, 259
Circuit, *see* Switching circuits
Classical algebra, 1
Closed switch, 19

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4

Copyright © 2004 John Wiley & Sons, Inc.

Closure of an operation, 3, 48
Code, 264
  BCH, 284
  binary, 266
  cyclic, 292
  error-correcting, 265
  error-detecting, 264
  linear, 276
  (n, k)-, 266
  parity check, 267, 274
  polynomial, 270
  repeating, 268
Code rate, 267
Code word, 266
Coefficient, 166
Colorings, 128
Common divisor, 184, 299
  multiple, 184, 303
Commutativity, 3, 14, 48, 148, 167
Commutator, 101
Complement, 8, 14
Complex numbers, 4, 223
Complex roots, 191
Composition of relations, 150
  of functions, 49
  of relations, 150
Conclusion, 293
Concatenation, 140
Conditional, 17
Cone, n-agonal, 116
Congruence, 77, 79, 145
  linear, 197
Congruence class, 77, 145
Conjugate, 191
Conjunctive normal form, 45
Constant polynomial, 167
Constructible number, 252
Constructible point, 252
Construction of polygons, 259
Contradiction, 17
  proof by, 294
Converse, 295
Convolution fraction, 175
Convolution of functions, 173
  of sequences, 168
Coset, 79, 82
Coset leader, 280
Countable, 221
Counterexample, 295
Cross-ratio, 101
Crystalline lattice, 120
Crystallographic group, 120
Cube, duplication of, 251, 256
Cube, rotation group of, 114
Cycle, 64
  disjoint, 66
Cyclic code, 291
Cyclic group, 56
Cyclic monoid, 139, 150
Cyclic subgroup, 57
Cyclotomic polynomial, 195
Cylinder, n-agonal, 116

Da Vinci, L., 109
Decode, 280
Dedekind cuts, 6
Degree, of an extension, 219
  of a polynomial, 166
Delta function, 172
De Morgan's laws, 15
Derivative, 292
Detecting errors, 280
Determinant, 110, 112, 287
Diagram, state, 143
  tree, 148
  Venn, 8, 32
Difference, 9
  symmetric, 9
Dihedral group, 58, 90
Dirac delta function, 172
Direct product, 2, 91, 164
Direct sum, 91
Disjoint cycles, 66
Disjunctive normal form, 30
Distributions, 175
Distributivity, 4, 14, 156
Division, 184, 299
Division algorithm, 180, 181, 298
Divisor of zero, 159
Dodecahedron, 112
  rotation group, 114
Domain, 172
  left Ore, 172
Domain, integral, 159
Duality in boolean algebras, 15
Duality of regular solids, 112
Duplication of the cube, 251, 256
Dürer, A., 246
DVD, 264

Eigenvalue, 108, 109, 111
Eigenvector, 108, 110
Eisenstein's criterion, 194
Element of a set, 7
Empty set, 7
Encode, 266, 280
Encoding matrix, 276
Endomorphism ring, 178

Equivalence class, 77
Equivalence, logical, 295
Equivalence relation, 5, 77
Equivalent circuits, 20
Error-correcting code, 264
Error-detecting code, 264
Error polynomial, 275
Euclidean algorithm, 185, 300
Euclidean group, 104
Euclidean ring, 181
Euclid's lemma, 302
Euclid's theorem, 298
Euler, L., 103, 236, 242, 260
Euler ϕ-function, 103
Even permutation, 68
Exclusive OR, 10, 29
Extension field, 218
  degree of, 219
  finite, 219

Factor, 184, *see also* Quotient
Factor theorem, 182
Faithful action of a group, 96
Faithful representation, 109
Feedback shift register, 231, 272
Fermat, P., 103, 260
Fermat primes, 260
Fermat's last theorem, 305
Field, 4, 160
  finite, 6, 45, 225
  Galois, 6, 227
  primitive element in, 229
  skew, 172
Field extension, 218
Field of, convolution fractions, 175
  fractions, 170
  quotients, 170
  rational functions, 172, 221
Fifteen puzzle, 74
Finite, abelian group, 92
  extension, 219
  field, 6, 45, 225
  geometry, 242
  group, 56
  groups in three dimensions, 116
  groups in two dimensions, 109
Finite-state machine, 142
First isomorphism theorem, 87, 210
Fixed points, 125
Flip-flop, 46
Formal power series, 169
Fractional differentiation and integration, 175
Fractions, field of, 170
  left, 172

Free, group, 75
  monoid, 140
  semigroup, 140
Full adder, 35
Frobenius automorphism, 235
Function, bijective, 50, 63
  composition of, 49
  delta, 172
  generalized, 175
  Heaviside, 175
  impulse, 172
  injective, 49
  inverse, 49
  one-to-one, 49
  onto, 50
  surjective, 49
  transition, 142
Fundamental theorem of, algebra, 6, 190
  arithmetic, 187

Galois, É., 6, 47, 225
Galois field, 6, 227
Galois theory, 47,
Gate, 36
Gauss, C. F., 6, 190, 260
Gaussian integers, 72, 183
Gauss' lemma, 193
Generalized function, 175
General linear group, 107
Generator, group, 56
  idempotent, 292
  matrix, 276
  of a monoid, 139
  polynomial, 270
Geometry, 242
Goldbach conjecture, 304
Graeco-Latin square, *see* Orthogonal latin squares
Greatest common divisor, 21, 184, 299
Greatest lower bound, 25
Group, 48
  abelian, 48
  alternating, 70, 82, 86, 88
  automorphism, 75
  center, 74
  circle, 88, 98
  commutative, 48
  crystallographic, 120
  cyclic, 56
  dihedral, 58, 90
  euclidean, 104
  factor, 83
  finite, 56
  free, 75
  general linear, 107

Group (*continued*)
generator, 56
icosahedral, 115
infinite, 56
isomorphism, 60
Klein, 52, 90
matrix, 107
metabelian, 102
morphism, 60
noncommutative, 59
octahedral, 114
of a polynomial, 80
of a rectangle, 53, 55
of a square, 89
of low order, 94
of prime order, 80
order, 56
orthogonal, 105
permutation, 50
quaternion, 73, 95, 107
quotient, 83
simple, 86
special orthogonal, 98, 108
special unitary, 108
sporadic, 86
symmetric, 50, 63
symmetries, 51
tetrahedral, 113
translation, 49, 104
trivial, 49
unitary, 108
Group acting on a set, 96
Group isomorphism, 60
Group morphism, 60

Half adder, 34
Hamming distance, 268
Heaviside, O., 172, 175
Homomorphism, *see* Morphism
Hypothesis, 293

Icosahedral group, 115
Icosahedron, 112
rotation group, 114
Ideal, 204
left, 217
maximal, 216
prime, 216
principal, 204
Ideal crystalline lattice, 120
Idempotent element, 179
generator, 292
laws, 15
Identity, 4, 48, 137
function, 49
If and only if, 18, 295
Image, 87
Implication, 17, 293
Improper rotation, 52, 58, 108
Impulse functions, 172
Inclusion, 7
Index of a subgroup, 80
Induction, 296
extended, 297
strong, 297
Infinite group, 56
Information rate, 267
Injective, 49, 86
Input values, 142
Integers, 5, 156, 296
gaussian, 72, 183
Integers modulo m, 78
Integral domain, 159
Interlacing shuffle, 74
International standard book number, 289
Intersection of sets, 8
Inverse, 3, 48, 188
function, 49
Inversion theorem, 50
Invertible element, 3, 48, 188
Irrational roots, 192
Irreducible, 189
polynomial, 190
Isometry, 51, 112
Isomorphism, 5
boolean algebra, 39
group, 60
monoid, 141
ring, 162
Isomorphism theorems for groups, 87, 101, 102
Isomorphism theorems for rings, 210, 215

Join, 14
Juxtaposition, 140

Kernel, 86
Klein, F., 52
Klein 4-group, 52, 90
Kronecker, L., 5

Lagrange, J., 79
Lagrange's theorem, 80
Latin square, 236
orthogonal, 239
Lattice, 25
crystalline, 120

Least common multiple, 21, 184, 303
Least upper bound, 25
Left coset, 82
Left ideal, 217
Linear code, 276
  cyclic, 291
Linear congruences, 197,
Linear transformation, 5, 104
Lines in a geometry, 242
Length of a vector, 105
Local ring, 216
Logically equivalent, 17, 295
Logic of propositions, 16

Machine, 142
  monoid of, 145
  parity checker, 143
  semigroup of, 142
Magic square, 247
Mathematical theory, 295
Matrix
  eigenvalue of, 108
  eigenvector of, 108
  encoding, 276
  generator, 276
  nonsingular, 4
  orthogonal, 105
  parity check, 278
  unitary, 108
Matrix group, 107
Matrix representation, 109
  of codes, 276
Matrix ring, 166
Maximal ideal, 216
Meet, 14
Metabelian group, 102
Metric, 291
Mikusinski, J., 173
Modern algebra, 2
Modular representation, 200
Modulo *m*, 78
Modulo 2 adder, 28
Monic polynomial, 234
Monoid, 137
  automorphism, 151
  cyclic, 139, 150
  free, 140
  generator, 139
  morphism, 141
  morphism theorem, 150
  quotient, 145
  representation theorem, 150
  transformations, 138
Monoid isomorphism, 141
Monoid morphism, 141
Monoid of a machine, 145
Monoid of transformations, 138
Morphism, 5
  boolean algebra, 39
  group, 60
  monoid, 141
  ring, 172
Morphism theorem for, groups, 87, 101, 102
  monoids, 150
  rings, 210
Mutually orthogonal squares, 239

NAND gate, 28, 36
Necklace problems, 126
Network, *see* Switching circuits
Nilpotent element, 216
Nonsingular matrix, 4
NOR gate, 28, 36
Normal form, conjunctive, 45
  disjunctive, 30
Normal subgroup, 82
NOT gate, 36

Octahedral group, 114
Octahedron, 112
  rotation group, 114
Odd permutation, 68
One-to-one, 49
One-to-one correspondence, 50
Onto, 50
Open switch, 19
Operation, 2
Operational calculus, 172
Operator, 175
Orbit, 65, 78, 97
Order of a group, 56
Order of an element, 56
OR gate, 36
Orthogonal group, 105
Orthogonal latin squares, 239
Orthogonal matrix, 105
Output, 142

Parallel circuit, 19
Parallelism, 242
Parity check, 278
  code, 267, 274
  machine, 143
  matrix, 278
Parity of a permutation, 69
Partial order, 24
Partition, 77
Pattern recognition, 147

Permutation, 50, 63
  even, 68
  odd, 68
  parity, 69
Permutation group, 50, 63
Phi function, 103
Plane, affine, 242
  projective, 245
Points of a geometry, 242
Pole of a rotation, 116
Pólya, G., 125, 134
Pólya-Burnside enumeration, 124
Polygons, construction, 259
Polyhedron, dual, 112
Polynomial, 166
  coefficients, 166
  constant, 167
  cyclotomic, 195
  degree, 166
  equality, 166
  group of symmetries, 74
  irreducible, 189, 190
  monic, 234
  primitive, 231, 275
  reducible, 190
  zero, 166
Polynomial code, 270
Polynomial equations, 1, 47
Polynomial representation of codes, 270
Polynomial ring, 167
Poset, 24
Positive integers, 3
Power series, 169
Power set, 8
Prime, 189, 294, 297
  factorization theorem, 21, 302
  Fermat, 260
  ideal, 216
Prime order group, 80
Primitive element, 229
Primitive polynomial, 231, 275
Principal ideal, 204
Principal ideal ring, 205
Product group, 91
Product ring, 164
Projective plane, 245
Projective space, 102
Proof, 293
  by contradiction, 294
  by reduction to cases, 293
  direct, 293
Proper rotation, 52, 58, 108
  symmetry, 52
Propositional calculus, 18

Quaternion, group, 73, 95, 107
  ring, 172, 177
Quotient, 181, 299
  field of, 170
  group, 83
  monoid, 145
  ring, 206
  set, 72
  structure, 5

Radical of a ring, 216
Rational functions, 172, 221
Rational number, 6, 48, 78, 170
Rational roots theorem, 192
Real numbers, 6, 48, 156
Real numbers modulo one, 88
Real projective space, 102
Rectangle, symmetries of, 53, 55
Reducible, 190
Redundant digits, 267
Reflexivity, 23, 77
Register, shift, 231, 272
Regular n-gon, rotations, 58
Regular polygons, construction, 259
Regular solid, 112
Relation, 76
  composition of, 150
  congruence, 77, 145
  equivalence, 77
  partial order, 24
Relatively prime, 301
Remainder, 181, 299
Remainder theorem, 182
Repeating code, 268
Representation, faithful, 109
  matrix, 109
  modular, 200
  residue, 200
Representation theorem for, boolean
  algebras, 39, 40
  groups, 71
  monoids, 150
Representative of a coset, 79
Residue representation, 200
Right coset, 79
Ring, 155
  boolean, 158, 176
  characteristic, 226
  commutative, 156
  endomorphism, 178
  euclidean, 181
  factor, 206
  field, 160
  ideal of, 204

integral domain, 159
local, 216
matrix, 166
morphism, 172
morphism theorem, 210
nontrivial, 159
polynomial, 166
principal ideal, 205
product, 164
quotient, 206
radical of, 216
simple, 217
subring of, 161
trivial, 159
Ring isomorphism, 162
Ring morphism, 162, 210
Ring of, formal power series, 169
matrices, 166
polynomials, 166
quaternions, 172, 177
sequences, 168
Roots, 183
complex, 190, 191
irrational, 191
rational, 192
Rotations, 52, 58, 108
Rotations of, a cube, 114
a dodecahedron, 114
an icosahedron, 114
an *n*-gon, 58
an octahedron, 114
a tetrahedron, 113
Ruler and compass, 251

Second isomorphism theorem, 102, 215
Semigroup, 137
free, 140
Semigroup of a machine, 145
Sequences, ring of, 168
Serial adder, 152
Series, power, 169
Series circuit, 19
Series-parallel circuit, 20
Set, 7
Sets, algebra of, 7
Shannon, C. E., 267
Shift register, 231, 272
Shuffle, interlacing, 74
Simple, group, 86
ring, 217
Simplification of circuits, 26
Skew field, 172
Smallest subfield, 220
Solids, Archimedean, 121
Solids, regular, 112
Space, projective, 102
Special orthogonal group, 98, 108
Special unitary group, 108
Sphere, 116
Sporadic groups, 86
Square, latin, 236
magic, 247
orthogonal latin, 239
Square-free integer, 22, 302
Squaring the circle, 251, 259
Stabilizer, 97
Standard basis, 104
Standard matrix, 105, 164
State, 142
diagram, 143
Step function, 175
Stone's representation theorem, 40
Straight-edge and compass constructions, 251
Structure, algebraic, 4
Structure, quotient, 5
Subfield, 218
smallest, 220
Subgroup, 54
commutator, 101
cyclic, 57
index of, 80
normal, 92
Submonoid, 150
Subring, 161
Subset, 7
Substructure, 4
Sum, direct, 91
Surjective, 49
Switch, 19
Switching circuits, 20
bridge, 43
number of, 130
n-variable, 28
series-parallel, 28
Switching function, 27
Sylow theorems, 85
Symmetric condition, 23, 77
Symmetric difference, 9, 28
Symmetric group, 50, 63
Symmetries of a
figure, 3, 51
polynomial, 74
rectangle, 53, 55
set, 50
square, 89
Symmetry, proper, 52
Syndrome, 280

Table, 3
  truth, 16
Tarry, G., 242
Tautology, 17
Tetrahedral group, 113
Tetrahedron, 112
  rotation group, 113
Third isomorphism theorem, 102, 215
Titchmarsh's theorem, 174
Transcendental, 221
Transformation monoid, 138
Transient conditions in circuits, 49
Transistor gates, 36
Transition function, 142
Transitivity, 23, 77
Translation, 49, 104
Transpose of a matrix, 104
Transposition, 68
Tree diagram, 148
Trisection of an angle, 251, 257
Trivial group, 49
Trivial ring, 159
Truth table, 16

Unary operation, 2, 8, 14
Underlying set, 4
Union of sets, 8
Unique factorization theorem, 189
Unitary group, 108
Unit element, 188. *See also* Identity; Invertible element
Unity, *see* Identity
Universal bounds, 25

Vandermonde determinant, 287
Vector space, 5
Venn diagram, 8, 32

Well ordering axiom, 296
Wilson's theorem, 202
Words, 140

Zero, 14, 155
Zero divisor, 159
Zero polynomial, 166

PURE AND APPLIED MATHEMATICS
A Wiley-Interscience Series of Texts, Monographs, and Tracts

Founded by RICHARD COURANT
Editors: MYRON B. ALLEN III, DAVID A. COX, PETER LAX
Editors Emeriti: PETER HILTON, HARRY HOCHSTADT, JOHN TOLAND

ADÁMEK, HERRLICH, and STRECKER—Abstract and Concrete Categories
ADAMOWICZ and ZBIERSKI—Logic of Mathematics
AINSWORTH and ODEN—A Posteriori Error Estimation in Finite Element Analysis
AKIVIS and GOLDBERG—Conformal Differential Geometry and Its Generalizations
ALLEN and ISAACSON—Numerical Analysis for Applied Science
*ARTIN—Geometric Algebra
AUBIN—Applied Functional Analysis, Second Edition
AZIZOV and IOKHVIDOV—Linear Operators in Spaces with an Indefinite Metric
BERG—The Fourier-Analytic Proof of Quadratic Reciprocity
BERMAN, NEUMANN, and STERN—Nonnegative Matrices in Dynamic Systems
BERKOVITZ—Convexity and Optimization in $\mathbb{R}^n$
BOYARINTSEV—Methods of Solving Singular Systems of Ordinary Differential Equations
BURK—Lebesgue Measure and Integration: An Introduction
*CARTER—Finite Groups of Lie Type
CASTILLO, COBO, JUBETE, and PRUNEDA—Orthogonal Sets and Polar Methods in Linear Algebra: Applications to Matrix Calculations, Systems of Equations, Inequalities, and Linear Programming
CASTILLO, CONEJO, PEDREGAL, GARCIÁ, and ALGUACIL—Building and Solving Mathematical Programming Models in Engineering and Science
CHATELIN—Eigenvalues of Matrices
CLARK—Mathematical Bioeconomics: The Optimal Management of Renewable Resources, Second Edition
†COX—Primes of the Form $x^2 + ny^2$: Fermat, Class Field Theory, and Complex Multiplication
*CURTIS and REINER—Representation Theory of Finite Groups and Associative Algebras
*CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume I
CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume II
DINCULEANU—Vector Integration and Stochastic Integration in Banach Spaces
*DUNFORD and SCHWARTZ—Linear Operators
Part 1—General Theory
Part 2—Spectral Theory, Self Adjoint Operators in Hilbert Space
Part 3—Spectral Operators

* Now available in a lower priced paperback edition in the Wiley Classics Library.
† Now available in paperback.

Modern Algebra with Applications, *Second Edition*, by William J. Gilbert and W. Keith Nicholson
ISBN 0-471-41451-4 Copyright © 2004 John Wiley & Sons, Inc.

FARINA and RINALDI—Positive Linear Systems: Theory and Applications
FOLLAND—Real Analysis: Modern Techniques and Their Applications
FRÖLICHER and KRIEGL—Linear Spaces and Differentiation Theory
GARDINER—Teichmüller Theory and Quadratic Differentials
GILBERT and NICHOLSON—Modern Algebra with Applications, Second Edition
GREENE and KRANTZ—Function Theory of One Complex Variable
*GRIFFITHS and HARRIS—Principles of Algebraic Geometry
GRILLET—Algebra
GROVE—Groups and Characters
GUSTAFSSON, KREISS and OLIGER—Time Dependent Problems and Difference Methods
HANNA and ROWLAND—Fourier Series, Transforms, and Boundary Value Problems, Second Edition
*HENRICI—Applied and Computational Complex Analysis
Volume 1, Power Series—Integration—Conformal Mapping—Location of Zeros
Volume 2, Special Functions—Integral Transforms—Asymptotics—Continued Fractions
Volume 3, Discrete Fourier Analysis, Cauchy Integrals, Construction of Conformal Maps, Univalent Functions
*HILTON and WU—A Course in Modern Algebra
*HOCHSTADT—Integral Equations
JOST—Two-Dimensional Geometric Variational Procedures
KHAMSI and KIRK—An Introduction to Metric Spaces and Fixed Point Theory
*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume I
*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume II
KOSHY—Fibonacci and Lucas Numbers with Applications
LAX—Functional Analysis
LAX—Linear Algebra
LOGAN—An Introduction to Nonlinear Partial Differential Equations
McCONNELL and ROBSON—Noncommutative Noetherian Rings
MORRISON—Functional Analysis: An Introduction to Banach Space Theory
NAYFEH—Perturbation Methods
NAYFEH and MOOK—Nonlinear Oscillations
PANDEY—The Hilbert Transform of Schwartz Distributions and Applications
PETKOV—Geometry of Reflecting Rays and Inverse Spectral Problems
*PRENTER—Splines and Variational Methods
RAO—Measure Theory and Integration
RASSIAS and SIMSA—Finite Sums Decompositions in Mathematical Analysis
RENELT—Elliptic Systems and Quasiconformal Mappings
RIVLIN—Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory, Second Edition
ROCKAFELLAR—Network Flows and Monotropic Optimization
ROITMAN—Introduction to Modern Set Theory
*RUDIN—Fourier Analysis on Groups
SENDOV—The Averaged Moduli of Smoothness: Applications in Numerical Methods and Approximations

* Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in paperback.

SENDOV and POPOV—The Averaged Moduli of Smoothness
*SIEGEL—Topics in Complex Function Theory
Volume 1—Elliptic Functions and Uniformization Theory
Volume 2—Automorphic Functions and Abelian Integrals
Volume 3—Abelian Functions and Modular Functions of Several Variables
SMITH and ROMANOWSKA—Post-Modern Algebra
STAKGOLD—Green's Functions and Boundary Value Problems, Second Edition
*STOKER—Differential Geometry
*STOKER—Nonlinear Vibrations in Mechanical and Electrical Systems
*STOKER—Water Waves: The Mathematical Theory with Applications
WATKINS—Fundamentals of Matrix Computations, Second Edition
WESSELING—An Introduction to Multigrid Methods
†WHITHAM—Linear and Nonlinear Waves
†ZAUDERER—Partial Differential Equations of Applied Mathematics, Second Edition

* Now available in a lower priced paperback edition in the Wiley Classics Library.
† Now available in paperback.